

格致方法 · 定量研究系列 吴晓刚 主编

功效分析概论：两组差异研究

[美] E.C.赫德伯格（E.C. Hedberg）著
洪岩璧 译

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

71

格致出版社 上海人民出版社

格致方法·定量研究系列　吴晓刚　主编

功效分析概论：两组差异研究

[美]E.C.赫德伯格（E.C.Hedberg）　著
洪岩璧　译

SAGE Publications, Inc.
格致出版社　上海人民出版社

图书在版编目(CIP)数据

功效分析概论:两组差异研究/(美)E.C.赫德伯格著;洪岩璧译.—上海:格致出版社:上海人民出版社,2021.11
(格致方法·定量研究系列)
ISBN 978-7-5432-3278-5

Ⅰ.①功… Ⅱ.①E… ②洪… Ⅲ.①统计分析-研究
Ⅳ.①C812

中国版本图书馆 CIP 数据核字(2021)第 185228 号

责任编辑 张苗凤

格致方法·定量研究系列
功效分析概论:两组差异研究
[美]E.C.赫德伯格 著
洪岩璧 译

出　版	格致出版社 上海人民出版社 (201101 上海市闵行区号景路 159 弄 C 座)
发　行	上海人民出版社发行中心
印　刷	浙江临安曙光印务有限公司
开　本	920×1168 1/32
印　张	7.25
字　数	144,000
版　次	2021 年 11 月第 1 版
印　次	2021 年 11 月第 1 次印刷

ISBN 978-7-5432-3278-5/C·257
定　价 48.00 元

出版说明

由吴晓刚(原香港科技大学教授,现任上海纽约大学教授)主编的"格致方法·定量研究系列"丛书,精选了世界著名的SAGE出版社定量社会科学研究丛书,翻译成中文,起初集结成八册,于2011年出版。这套丛书自出版以来,受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择,该丛书经过修订和校正,于2012年以单行本的形式再次出版发行,共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化,我们又从丛书中精选了三十多个品种,译成中文,以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复,但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始,到多元线性回归模型结束,内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型,以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法,将重点放在因变量为定类数据的分析模型上,包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型,包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生,一直鼓励和支持我将两门课的讲稿结集出版,并帮助我将原来的英文课程讲稿译成了中文。但是,由于种种原因,这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩,每本书都写得通俗易懂,与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译,以飨中文读者时,我非常支持这个想法,因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办"应用社会科学研究方法研修班"，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

我很荣幸向大家介绍 E. C. 赫德伯格（E. C. Hedberg）的《功效分析概论：两组差异研究》一书。统计功效是指假定零假设为假、备择假设为真的情况下，统计检验显著的概率。这是研究设计的关键组成部分，需要由伦理委员会、出资机构和出版方进行审查和评估。伦理委员会希望了解该研究是否能充分证明对参与者的风险。出资机构同样关注该研究设计（包括统计功效）是否满足科学性目标的要求。此外，给定数据收集的成本，出资方也关注成本效益问题，尤其是样本量不能超过所需规模。最后，期刊和其他学术出版机构也关注统计功效，因为这对稳健性、结果的信度和可复制性皆至关重要。

本书通过干预组和控制组两组间的比较介绍了统计功效，可作为研究生定量方法课程的补充学习材料。对于学生、教师和其他有志于学习统计功效的研究者而言，本书也颇有价值。本书内容全面，提供了所有的相关信息，包括对主要统计分布、假设检验和错误类别的梳理。本书组织明晰、结构紧凑。

赫德伯格为读者打基础的策略是，先通过详细的简单案例传递要点，然后拓展到更复杂但简略的案例，为读者思考如何运用做准备。在介绍功效分析是什么、为什么需要、何时使用(第 1 章)，并对主要的分布进行简明的回顾(第 2 章)之后，作者开始处理总体标准差已知时假设检验和功效分析中的议题(第 3 章)，以及需要估计总体标准差时的相关议题(第 4 章)。接下来，作者讨论检验均衡设计情况下的组间均值差异时包含协变量的情况(第 5 章)、二层聚合随机试验(第 6 章)，以及二层多点随机试验(第 7 章)。功效分析实例使用主流的软件包(SPSS、Stata 和 R)，完整的编码和输出可参见网络资源：study.sagepub.com/hedberg。

随后的两章转向功效分析的实践操作，这对实践研究者尤其有用。第 8 章梳理了进行功效分析所需的假定。赫德伯格对这一章进行了如下总结："虽然研究者永远无法确定地预测未来数据的参数……但臆测相关的假定或用传统观点进行臆测绝非好主意。"赫德伯格为如何把前人的研究用于功效分析假定提供了专门的指南，包括那些并非完全契合当下所用变量的前人的研究。第 9 章为如何以简洁完备的形式报告功效分析提出了建议。当研究者需要向伦理委员会、出资机构和同行评审期刊解释研究时，这是必备技能。本书结尾部分简述了一些更进阶的议题，并给出了相关文献，以便有兴趣的读者进一步学习。

如今，社会科学和行为科学对由行政记录、交易数据、网上活动、社交媒体互动以及 GPS 定位信息等大规模数据的兴趣与日俱增。当数据量非常大时，统计功效很少被关注(虽然也有其他的研究设计视角挑战这些结果的效度)。然而，

即使在“大数据”时代，我们仍然需要小样本研究，对于这些研究而言，本书所提供的内容非常重要。我很享受阅读本书，希望你也会如此。

芭芭拉·恩特威斯尔

前 言

倘若某个研究领域依赖统计推断作为其证据基础，那么其结果的信度，无论是否显著，很大程度上都取决于统计功效。当统计功效很低时，知识会有所损失，随机噪声会取而代之成为知识。当然，这对"功效很重要"这个简单道理进行了过度简化。

本书是功效分析的概论。我并非随意取了"概论"这个书名，而确实是仅仅概述了功效分析的核心要素。我以两组均值的比较为例，并基于这一研究问题，和读者一起从简单走向复杂的思考。从协变量到聚类效应，两组比较有时会变成一项复杂任务，功效分析也会相应变得繁杂。

通过这本概论，我希望读者能洞悉他们自己所做的分析，并据此奠定基础，以更好地理解他们应用于研究的功效分析。

至于细节，我尽量保持符号标记的前后一致。统计学是一种抽象语言，每位教师都有自己的语调。因此，虽然我难以使我的多层模型符号同时与劳登布什* 和赫奇斯** 的保持

---

* 斯蒂芬・劳登布什(Stephen Raudenbush)是美国社会统计学家，著有《分层线性模型：应用与数据分析方法》。——译者注

** 拉里・V.赫奇斯(Larry V.Hedges)是美国知名的教育统计和评估学者，研究领域横跨统计学、社会学、心理学和教育政策。——译者注

一致，但我至少尽量保持我自己著作的内部一致。这意味着我的符号标记会与他人的符号有所不同。我尝试注明与传统标记的主要不同之处。简言之，这是一个需要关注的问题，譬如我的 τ 和劳登布什的 τ 含义就不同。

本书使用了大量实例。其中一些实例的分析使用了真实数据。读者需注意，这些例子中使用的数据都是出于教学目的。虽然数据是真实的，但它们都抽取自更大的数据库，仅用于说明功效分析。因此，分析结果不能作为经验证据来加以引用。我很感激那些把数据放在网络资源库共享的研究者，以供像我这样的人在夜阑人静之时筛选。

末了，我想明确感谢他人的知识贡献。R 软件的“texreg”包（Leifeld, 2013）、R 软件的“xtable”包（Dahl, 2009）和用于画图的 R 软件包“tikzDevice”（Sharpsteen & Bracken, 2013）使本书在 LaTeX 系统*中的写作和排版变得更为便捷。

我希望你能喜欢这本书，从中获取统计功效的些许知识。功效分析实例使用主流的软件包（SPSS、Stata 和 R），完整的编码和输出可参见网络资源：study.sagepub.com/hedberg。

---

* LaTeX 是一种基于 TEX 的排版系统，由美国计算机学家莱斯利·兰伯特（Leslie Lamport）在 20 世纪 80 年代初期开发。使用者利用这种格式，即使没有排版和程序设计的知识，也可以充分发挥由 TeX 所提供的强大功能。LaTeX 尤其适用于生成复杂表格和数学公式，因此非常适合用于生成高印刷质量的科技和数学类文档。详细信息参见 https://www.latex-project.org。——译者注

致谢

本研究得到了美国教育署教育科学研究所(项目号:R305D140019)和芝加哥大学全国民意调查中心的支持。本书观点皆来自作者本人,并不代表美国教育署教育科学研究所的看法。

作者感谢下述人员对本书写作给予的支持和建议:

拉里·赫奇斯,西北大学;

查尔斯·卡茨(Charles Katz),亚利桑那州立大学;

阿伦·凯珀(Arend Kuyper),西北大学;

丹妮尔·华莱士(Danielle Wallace),亚利桑那州立大学。

作者还要感谢下述评审专家的贡献:

克里斯·阿伯森(Chris Aberson),洪堡州立大学;

莱斯莉·埃科尔斯(Leslie Echols),密苏里州立大学;

埃琳·M.费克特(Erin M.Fekete),印第安纳大学;

斯蒂芬妮·J.琼斯(Stephanie J.Jones),得克萨斯理工大学;

卡琳·林德斯特伦·布雷默(Karin Lindstrom Bremer),明尼苏达州立大学曼卡托(Mankato)分校;

弗雷德·奥斯瓦尔德(Fred Oswald),赖斯大学;

贾森·波潘(Jason Popan),得克萨斯大学泛美分校;

加里·波波利(Gary Popoli),史蒂文森大学;

本·凯尔西(Ben Kelcey),辛辛那提大学;

布赖恩·J.鲁尼(Bryan J.Rooney),康考迪亚大学埃德蒙顿分校。

Introduction to Power Analysis: Two-Group Studies
by E. C. Hedberg
English language editions published by SAGE Publications of Thousand Oaks, London, New Delhi, Singapore and Washington D.C., © 2018 by SAGE Publications, Inc.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2021.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。
上海市版权局著作权合同登记号:图字 09-2021-0463

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用 logistic 回归分析（第二版）
14. logit 与 probit：次序模型和多类别模型
15. 定序因变量的 logistic 回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic 回归入门
39. 解释概率模型：Logit、Probit 以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL 导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic 回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库
59. 应用人口学概论：数据来源与估计技术
60. 多元广义线性模型
61. 时间序列分析：回归技术（第二版）
62. 事件史和生存分析（第二版）
63. 样条回归模型
64. 定序题项回答理论：莫坎量表分析
65. LISREL 方法：多元回归中的交互作用
66. 蒙特卡罗模拟
67. 潜类别分析
68. 内容分析法导论（第二版）
69. 贝叶斯统计推断
70. 因素调查实验
71. 功效分析概论：两组差异研究
72. 多层结构方程模型

目 录

第1章

什么是功效分析？为何以及何时使用？

第1节 | 什么是统计功效?

统计功效是指给定条件下某个检验统计显著的概率,也即零假设事实上为假、备择假设为真的概率,这些给定条件包括可接受的不确定性、效应大小、抽样设计和样本量(Cohen, 1988)。功效水平接近0,表明检测到影响效应的几率很低;而功效水平接近1,则表明检测到效应的概率很高。社会科学的传统是研究设计的功效至少要达到0.8,即有80%的几率能检测到效应。功效与统计检验中不同类别的错误和某研究领域中知识的发展潜力直接相关。

研究中的错误

相对于某个零假设(比如锻炼不会影响体重),任何备择假设(比如锻炼能减肥)都有四种可能结果。第一种结果是研究者认为备择假设更有可能,而事实上备择假设确实为真。这一般被称为"显著"结果。譬如,研究者得出结论认为某个新课程能有效提升数学成绩,而事实上该课程确实提升了成绩。

第二种结果是研究者得出结论认为零假设更可能是真的,而事实上零假设为真。这一般被称作"虚无"结果。譬

如，研究者得出结论认为某种新药对头痛没有效果，而事实上该药确实对头痛无效。上述两种“显著”和“虚无”结果非常理想，因为研究者的推断和事实相符。

接下来的两种研究结果则属于错误(error)。首先是第一类错误，指研究者拒绝了零假设，倾向于接受备择假设，但事实上零假设为真。比如，研究者的结论是某种治疗对犯罪行为有影响，而事实上这一治疗对行为没有影响，那么就发生了第一类错误。传统上，研究中可以接受这种错误情况发生的比例是5%，对应的第一类错误是0.05。我们用希腊字母α来表示第一类错误，检验所对应的第一类错误水平是$\alpha=0.05$。

其次是第二类错误，此时研究者接受零假设，但事实上备择假设为真。譬如，结论认为某项政策对穷人的经济后果没有影响，而事实上该项政策影响了经济后果，此时就发生了第二类错误。传统上希望研究中发生此类情况的几率不高于20%，即对应的第二类错误水平为0.2。我们用希腊字母β表示第二类错误，检验所对应的第二类错误水平是$\beta=0.2$。某个检验的功效等于$1-\beta$，即当备择假设事实上为真时，该检验检测到这一效应的几率。

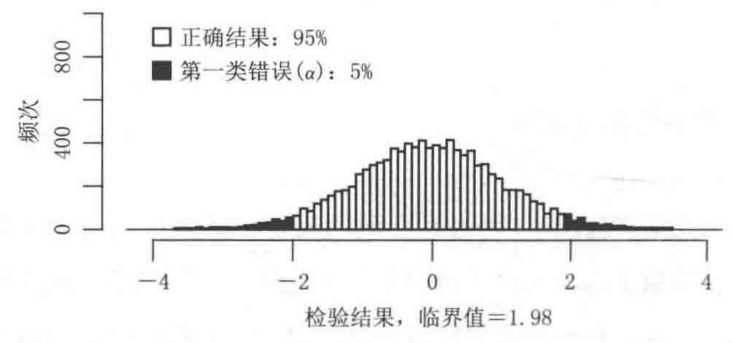
一个小模拟实验

在某些情况下，研究者会使用蒙特卡罗模拟法来确定功效。模拟法听起来很高深，但其前提设定非常简单。利用大多数统计软件包中都有的随机数生成器，研究者就能进行多种假定的编程。譬如，研究者可以假定变量服从均值为0、标准差为1的正态分布，把这两个数输入到随机数生成器中，

就能产生一组个案（也在研究者的掌控之下）。一旦产生了随机数字（即随机观测），研究者就能直接对模拟数据进行常规分析（软件并不会介意数据是假的），并记录所得的检验统计量和其他感兴趣的参数。这一过程需重复几千遍，每次模拟结果会略微有所不同。

这些模拟得到的结果可以描绘成直方图，以揭示模拟的抽样分布，这是一个检验中心极限定理的好方法。就功效分析而言，研究者也可以记录结果是统计显著的模拟次数。而显著结果的比例就是产生这一模拟结果的某组假定（即检验）的功效。

比如说，我们可以用下述假设检验的模拟来阐述两类错误（对假设检验的全面考察将贯穿本书）。第一个模拟产生各包括 50 个个案的两组数据，每组数据都来自均值为 0、标准差为 1 的正态分布。由于两组数据来自同一个分布，即均值为 0、标准差为 1 的标准正态分布，所以我们知道这两个组真实的均值差就是 0。一旦产生了这两组数据，就可以运用简单双尾 t 检验来确定这两个组的均值是否相等。我



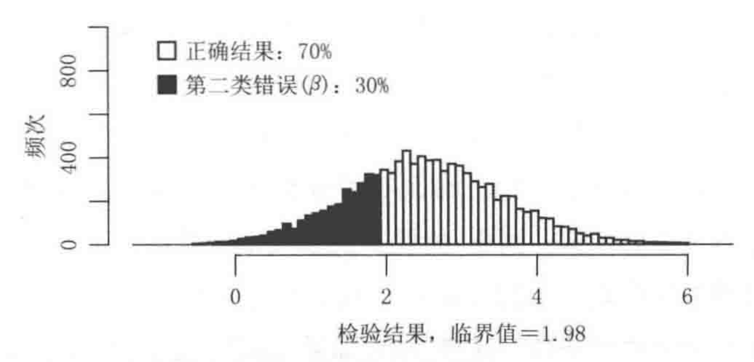
注：不正确的结果（即第一类错误）涂成了黑色。

图 1.1 样本量为 100、真实组间差为 0 个标准差（即零假设为真）的 10 000 次模拟结果双尾检验（$\alpha=0.05$）的直方图

们让这个运算运行 10 000 次，并在数据库中记录下每次结果的 t 检验。图 1.1 呈现了这些模拟结果的直方图。当我们设定 $\alpha=0.05$，则这一检验的临界值约为 -1.98 和 1.98。表 1.1 检验结果中大约 5%出现在临界值两端区域内，这对应于第一类错误发生的比率 $\alpha=0.05$。

接下来考察另一个模拟过程。在这个模拟中，我们从均值为 0、标准差为 1 的分布中抽取一组数据，从均值为 0.5、标准差为 1 的分布中抽取另一组数据。因此，我们知道真实的组间均值差异为 0.5 个标准差。我们还是进行 10 000 次模拟运算，并像之前那样记录 t 检验结果。图 1.2 呈现了这些模拟结果的直方图。同样，当我们设定 $\alpha=0.05$，这一检验的临界值约为 -1.98 和 1.98。图 1.2 表明大约 30%的结果处于两个临界值之间的区域，对应于第二类错误 $\beta=0.3$，换言之，统计功效约为 0.7。如果我们改变模拟的参数，比如样本量或效应值，那么模拟得到的结果也会有所不同。

在收集数据和未进行模拟之前，本书会运用假定和公式来先验地确定检验功效。



注：不正确的结果(即第二类错误)涂成了黑色。

图 1.2　样本量为 100、真实组间差为 0.5 个标准差(即备择假设为真)的 10 000 次模拟结果双尾检验的直方图

第2节 | 为什么做研究设计时需要考虑功效?

为什么在收集数据之前理解检验的功效非常重要? 理由有两点。功效分析很重要的第一个理由很简单:理性的研究者应当希望最大化检测到研究假设效应的几率,同时最小化研究参与者面临的潜在风险。这一理由来自经济和伦理两方面的考量。经济考量在于研究需要花钱,而研究结果若不具有结论性则等同于徒费资源。伦理考量关注如果干预给参与者带来风险,那么该研究的样本量应当仅限于为了获得所想要的功效而必需的最低规模。

因此,对能检测出某项实验效果的几率的先验估计,可以为项目的成功可能性和风险提供有用信息。如果功效分析表明成功的几率很低,就给研究团队提供了一个机会重新评估他们的计划,并对研究设计和抽样进行修改。如果功效很高,且干预风险较大,那么功效分析就能为缩减干预组的样本量提供证据。同样,假设不会发现某个效应的研究者也必须关注功效。他们必须对研究中非常微小的有意义的效应给予足够的功效。这样一来,如果没有发现统计上显著的差异,研究者才有足够信心说结果真的是来自实际中的无效应,而非由于研究的功效不足。

第二个理由更为复杂一些，但也同样重要。近来，一些出版物已经概述了一个令人困扰的现象，那就是很多已经发表在科学类文献中的研究发现不是真实的（比如 Ioannidis, 2005），而且很多发现无法被复制（如 Open Science Collaboration, 2015）。这一现象有诸多原因，其中大部分原因都已超出了本书论述的范围，但其中一个原因便是很多研究的统计功效过低。

再回顾一下任何检验的四种结果：显著、虚无、第一类错误和第二类错误。我们运用约安尼季斯（Ioannidis, 2005）提出的公式并设置一些假定，可以估计研究结果落入四种结果中每一类的比例。在下述公式中，R 是真实的被检验假设与不实的被检验假设的比值，α 是第一类错误的比率，β 是第二类错误的比率。用约安尼季斯（Ioannidis, 2005）文章中最简单表格的形式，"显著"发现的预期比例是：

$$\text{显著} = 100 \times \frac{(1-\beta)R}{R+1} \qquad [1.1]$$

"虚无"发现的比例是：

$$\text{虚无} = 100 \times \frac{(1-\alpha)}{R+1} \qquad [1.2]$$

属于第一类错误的发现的比例是：

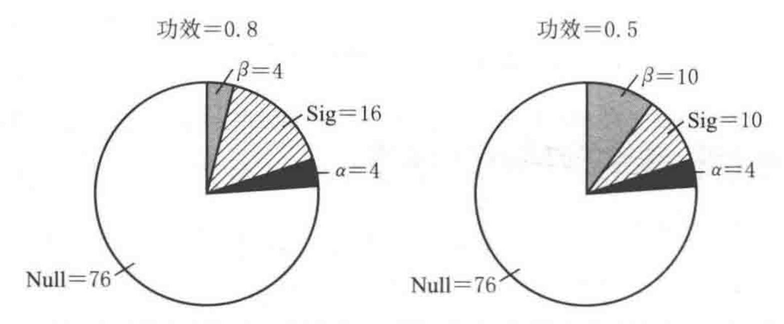
$$\text{第一类错误} = 100 \times \frac{\alpha}{R+1} \qquad [1.3]$$

最后，属于第二类错误的发现的比例是：

$$\text{第二类错误} = 100 \times \frac{\beta R}{R+1} \qquad [1.4]$$

假定所有被检验的假设中，20%实际上是真的。这意味着真实假设与不实假设的比值 $R=0.2/0.8=0.25$。接下来，假定所有假设都在 $\alpha=0.05$ 的水平上进行检验。这意味着虚无假设的比例大约为 $100\times[(1-0.05)/(0.25+1)]=76\%$。这也意味着第一类错误的比例约为 $100\times[0.05/(0.25+1)]=4\%$。因此，总的不实假设比例达到 80%，或者反过来说，如前文所言，真实假设的数量占 20%。图 1.3 用饼图呈现了这种 80%的情况，也标注了虚无假设和显著假设所占的部分。

为了说明功效的含义，首先假定仅当发现显著时研究结果才能被发表，如伊斯特布鲁克、高普兰、伯林和马修斯 (Easterbrook, Gopalan, Berlin, & Matthews, 1991)的研究就表明，统计显著的发现更可能被发表。其次，假定所有研究的功效是 0.8($\beta=0.2$，参见图 1.3 中的左图)。这意味着所有的假设中，第二类错误占比 $100\times[(0.2\times0.25)/(0.25+1)]=4\%$，所有假设中检验显著的占比 $100\times\{[(1-0.2)\times0.25]/[0.25+1]\}=16\%$。这说明 1/5 的真实假设未被发表，而已发表的假设中有 1/5 实际上是不实假设。



注：这里假定所检验假设中有 20%是真实的，检验的显著性水平 $\alpha=0.05$。

图 1.3 功效水平分别为 0.8 和 0.5 时，研究假设属于显著(Sig)、第一类错误(α)、虚无(Null)和第二类错误(β)四种情况的比例

当检验功效更低时，情况就更糟了。比如说，当功效为0.5（$\beta=0.5$，参见图1.3中的右图）。虚无假设不受影响，但是所有假设中属于第二类错误的比例为10%，而所有假设中真实显著的比例也是10%。这意味着14%的假设被发表了，而这些已发表的假设中几乎30%属于不实假设，而在更高功效情况下，不实假设比例只有20%。*

这一演示旨在说明随着特定研究领域中功效的下降，已发表结果中取决于偶然（即实际上为虚无）的比例会增加。已有研究表明某些领域的功效经常偏低（如Spybrook，2007），这也是一些研究者怀疑当前知识生产的原因之一。要想增加对知识的信心，办法之一就是提升建构这类知识的研究之功效。

---

* 4/(4+10)=28.57%，接近30%。高功效是指上文功效为0.8（$\beta=0.2$）的情况，如图1.3中的左图，4/(4+16)=20%。——译者注

第3节 | 何时应当进行功效分析？

简单来说，数据收集之前进行功效分析是最有用的。一旦数据收集完成，功效分析就像是对不显著结果进行"事后析误"（如果结果显著，那说明功效显然是足够的）。然而，对观测性数据进行功效分析仍可以获得有益的洞见。

实验

功效分析对随机试验尤其有用，也更简单。我们在后续章节中会看到，核心自变量（如属于干预组成员）和其他协变量之间的相关会使统计检验的先验期望更为复杂。这是因为必须考虑所有协变量之间相互关联的信息。当对核心自变量进行随机化，就可以（期望）消除这些相关，从而可以进行相对简单的功效分析。

观察性研究

在观察性研究中，通常在数据收集之前进行的是精度分析（precision analysis）而非功效分析。精度分析类似于功效分析，但不同于功效分析估计检测到显著效应的似然值（比

如两组间的均值差异)，对于给定样本量和设计的某总体均值或总体比例的可能性（对样本量规模确定的讨论参见 Lohr, 2009)，精度分析确定精度（比如误差幅度）。同样，在本书所讨论的不同功效分析中，研究者也会发现特定置信水平下达到某种精度所必需的样本量。这类分析对于旨在进行单变量分析的调查而言尤为重要，譬如民意测验。

再强调一遍，一旦数据已经收集完成，功效分析和精度分析就没太大用处了。如果某结果统计显著，即使功效分析显示较低的功效也不会使结果无效。如果结果不显著，有时候有助于确定在特定效应值下为了获得显著性结果，需要多大的样本。如果这一假设的样本量接近观测样本量，那么一个可信的推论应该是研究效应可能是存在的，只是这个研究不太走运。另一方面，若观测样本量已经远大于所需样本量，那么所能得出的结论就是并不存在有意义的效应。

第 4 节 | 显著性和效应

功效是关于统计显著性的。然而,功效分析的核心要素涉及效应值(可以被理解为实际的显著性)和样本量之间的相互作用。多年来,很多科学领域都局限于关注论述某些效应可能非零(nonzero)的能力,以及考察效应幅度的代价(参见 Ziliak & McCloskey, 2008)。这些是不同的度量。

阅读本书你会发现,一个足够大的样本甚至能够以某种精度检测到非常小的效应(即统计显著性)。但这可能意味着这项研究其实一无所获:它仅对无关紧要之事进行了精确估计。我们鼓励读者在进行功效分析时牢记这一点,因为研究的目的永远应当是发现(或否证)某种有意义且具有实际显著性的效应。

第 5 节 | 进行功效分析前需要了解什么？

不同检验的功效分析需要不同的要素。本书聚焦于运用线性回归考察两组间均值的差异。这一个案分析侧重于理解检验的组成要素：哪些参数构成这些要素、哪些可以假定，以及哪些在研究者的掌控之中。一般而言，任何参数检验的功效分析都需要理解下述六个方面：

（1）如何估计总体参数（比如组均值的差异）；

（2）如何计算总体参数的抽样方差（其平方根即为标准误）；

（3）如何进行检验计算；

（4）检验的哪个部分可以转化为无标尺（scale-free）的参数；

（5）如何计算零假设和备择假设的抽样分布曲线下的面积（一般利用计算机）；

（6）如何重新整理检验统计量，以剔除影响功效的其他因素，比如效应值或样本量。

本书是理解上述六个议题的指南，而以考察组均值差异为实例。我们希望读者通过这些实例操作，可以在自己的研究中探索如何进行并理解功效分析。

第6节 | 本书结构

下一章将简要回顾假设检验中用到的主要统计分布(标准正态分布、t 分布、χ^2 卡方分布和 F 分布)与标准正态分布的关系,以及通过正态分布了解这些分布相互间的关系。

第3章将聚焦于假设检验和不同类别错误的基本议题,以及在检验两组间均值差异中它们与简单功效分析的关系。然后本书的主要篇幅侧重讨论简单随机样本中两组比较的问题、涉及协变量的情况,以及更复杂样本中的情况。

本书结尾部分将讨论如何收集信息以设定功效分析的假定,以及如何报告功效分析的结果。结论章节提供了一些进阶功效分析的步骤和实例。

第 7 节 | 小结

在这个导论性章节中，我们探索了功效的一般性定义。通过模拟，我们近似计算了已知结果的检验的一个抽样分布，展示了组均值差异检验的第一类错误和第二类错误。

接下来我们简述了功效分析非常重要的理由。对于任何想要确证大多数已发表结果确实为真的学科而言，功效分析至关重要。虽然很多领域中已发表的结果可能实际上并不为真。

本章最后对本书剩余部分内容进行了简要介绍。

第2章

统计分布

本章简要讨论主要统计分布的基础性内容。本章的目的不是对个别分布进行推导,而是向读者展示主要统计分布与标准正态分布的关系。

理解分布之间的关系,对于我们认识在本书中发挥关键作用的 t 分布和 F 分布至关重要。在后续章节中,我们需要处理更为复杂的样本,要求运用双因素方差分析(ANOVA)模型(其中一个因素就是组别样本*)。通过本章了解分布之间的关系,有助于我们把复杂样本的 F 检验简化为 t 检验。

* 组别样本指不同的实验组构成一个样本。——译者注

第1节 | 正态分布的随机变量

我们首先讨论正态分布。中心极限定理(Rice, 2006)表明很多统计抽样分布都是正态分布,或者与正态分布存在某种关联。这就是为什么概论性课程往往非常重视“正态”曲线。如你在第1章的模拟中所见,检验统计量直方图呈现“钟形”(如图1.1)。

虽然数据集所包含的很多变量都接近正态分布,比如身高或体重,但这里的重点是样本统计量的分布。换言之,这一变量分布并非数据集中的某一列,而是由很多样本形成的统计量的假设分布。这个样本结果的分布通常被称为“抽样分布”。

基于某数据中每个组别的观测数据,变量Y的组间均值差异是一个随机变量,因为样本是随机抽取的(Lohr, 2009)。如果我们收集另一个随机样本,观察到的均值差异结果会有所不同。再收集第三个样本,就会有第三种结果,如此至于无穷。这就使得我们所感兴趣的组间差异统计量也是一个随机变量。

就像任何连续性随机变量那样,我们可以用密度函数来描述分布的形状。正态分布的密度函数$f(z)$如下(Rice, 2006):

$$f(z)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(z-\mu)^2/(2\sigma^2)} \qquad [2.1]$$

该函数的关键参数是分布均值 μ 和标准差 σ。如果变量 X 服从正态分布，我们就写作 $X \sim N(\mu;\ \sigma)$。因此，针对数据线 z 上的每一个点，我们都可以运用方程 2.1 找出密度值。图 2.1 描绘了方程 2.1 中 $\mu=0$ 和 $\sigma=1$ 的情况，这也被称作标准正态分布 $Z \sim N(0,\ 1)$。

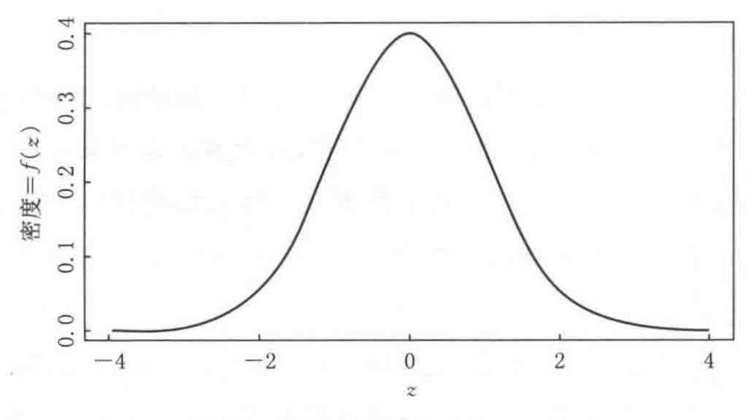


图 2.1 标准正态分布($\mu=0$, $\sigma=1$)

密度函数有一个诱人的特征，即曲线下涵盖的总面积等于 1。这使我们可以用下述方式来分割这一分布："z 之前的取值所含份额是……"，"z 之后的取值所含份额是……"，或"z_a 和 z_b 之间取值所含份额是……"。我们可以用正态分布的累积密度函数(CDF，标记为 Φ)进行这类计算，通常会用到统计表格或计算机。

譬如，若要计算总体中 X 小于或等于 x 的概率，已知总体均值为 μ，标准差为 σ，我们可以用累积密度函数进行下述表达：

$$P(X \leqslant x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad [2.2]$$

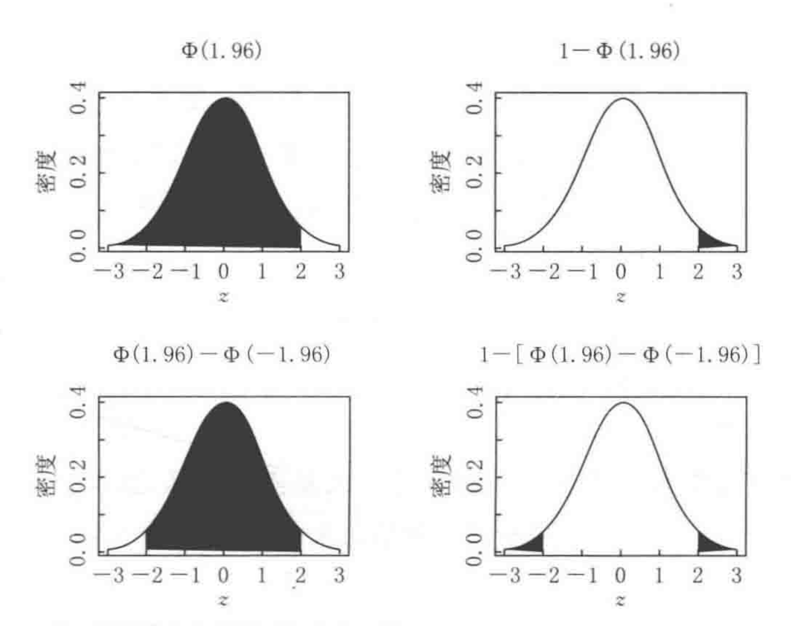
若要计算总体中 X 大于或等于 x 的概率，我们可以用累积密度函数进行如下表述：

$$P(X \geqslant x) = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad [2.3]$$

若要计算总体落入 x_a 和 x_b 之间的概率（假定 $x_a < x_b$），我们仍可以用累积密度函数进行如下表述：

$$P(x_a \leqslant X \leqslant x_b) = \Phi\left(\frac{x_b-\mu}{\sigma}\right) - \Phi\left(\frac{x_a-\mu}{\sigma}\right) \qquad [2.4]$$

其中，$\frac{x-\mu}{\sigma}$ 是 z 分数的通用计算式。多数统计包都有累积正态分布的函数，而功效分析则能充分利用它们。譬如，$\Phi(1.96)=0.975$，$1-\Phi(1.96)=0.025$，$\Phi(1.96)-\Phi(-1.96)=0.950$，以及 $1-[\Phi(1.96)-\Phi(-1.96)]=0.05$（参见图 2.2）。



注：阴影部分为累积正态分布函数 Φ。

图 2.2　正态分布面积

服从正态分布的抽样分布拥有相同的特征。抽样分布也有均值(通常被称作"期望"统计量)和标准差,抽样分布的标准差我们称作"标准误"。在下一章中,我们会探索如何利用正态分布,根据给定随机样本的统计量及其标准误来推论更大总体的情况。

第2节 | 卡方(χ^2)分布

很多情况下，统计分析会用到 χ^2 分布。χ^2 分布来自标准正态分布。虽然本书并不会直接用到 χ^2 分布，但它是通向 t 分布和 F 分布的重要连接点。该分布在检验方差与平方和时经常用到，因为方差与平方和都服从 χ^2 分布。[1]

本质上，χ^2 分布是相互独立的 Z 分布平方之和。

$$\sum_{i=1}^{v} Z_i^2 = \chi_v^2 \qquad [2.5]$$

在上述表达式中，v 是变量"自由度"的取值。[2]在这里，其取值等于我们相加的随机 Z 变量的个数。* 例如，当 $\alpha=0.05$ 时，我们经常使用 1.96 作为双尾 z 检验的临界值。一个 Z 变量的平方是一个自由度为 1 的 χ^2。翻检任何一本统计概论著作，你会发现自由度为 1 的 χ^2 临界值是 3.84，正好是 1.96 的平方。

因为 χ^2 由 Z 变量的平方和组成，所以 χ^2 的值始终为正。因此，其取值范围为 0 到任何一个正数。随着自由度(v)的增加，χ^2 分布越来越接近正态分布。

图 2.3 展现了不同密度的 χ^2 分布。和正态分布一样，我

---

* Z 变量指服从正态分布的变量。——译者注

们可以用累积密度函数来计算累积面积取值,比如分布的95%。这就使我们可以检验服从 χ^2 分布的随机变量是否很可能拒绝零假设。

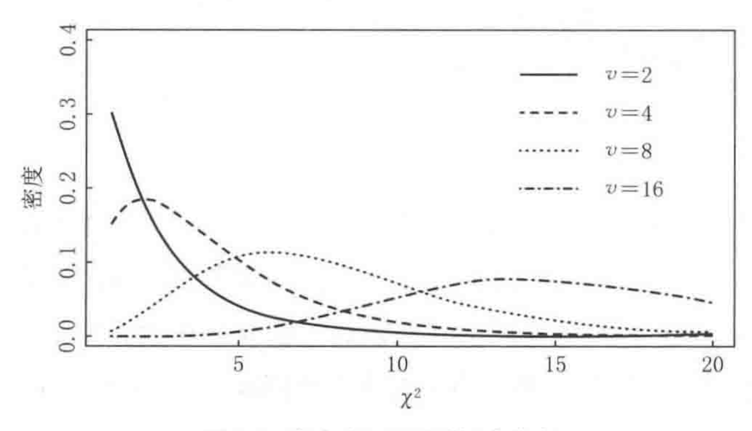


图 2.3 自由度(v)不同的 χ^2 分布

第3节 | t 分布

t 分布是本书最重要的关注点。后面我们会看到，除了 t 分布取决于检验的自由度(v)之外，t 分布和标准正态分布非常像。t 分布的公式和推导已超出本书的论述范畴。自由度为 v 的 t 分布与 Z 分布的关系如下(Rice，2006)：

$$t=\frac{Z}{\sqrt{\chi_v^2/v}}=\frac{Z}{\sqrt{\frac{1}{v}\sum_{i=1}^{v}Z_i^2}} \qquad [2.6]$$

t 分布比标准正态分布更“平”一些(参见图 2.4)。随着自由度(v)的增加，t 分布趋近于正态分布的形状。结果是，t 分布上的累积比例取值不同于标准正态分布上的累积比例取值。

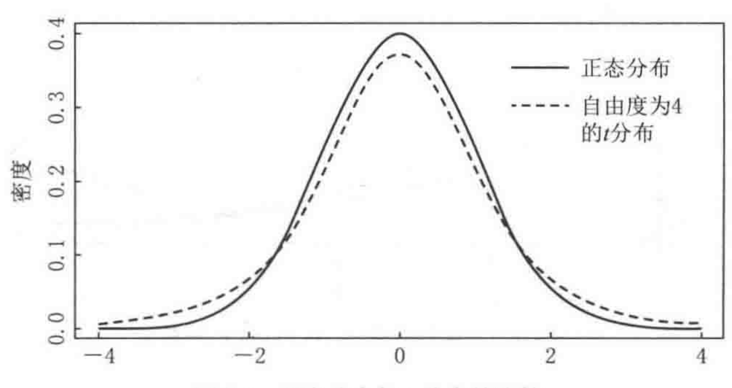


图 2.4 正态分布与 t 分布的比较

第 4 节 | *F* 分布

F 分布常用于方差比值。[3]因此,*F* 分布是基于多个 χ^2 形成的。一个随机 *F* 变量是两个 χ^2 分别除以各自的自由度(v)之后的比值:

$$F=\frac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2} \qquad [2.7]$$

因此,就像 *t* 分布那样,*F* 分布也是基于自由度的。*F* 分布取决于两个自由度,分别在分子和分母中(v_1 和 v_2)。图 2.5 提供了一些密度曲线的实例。

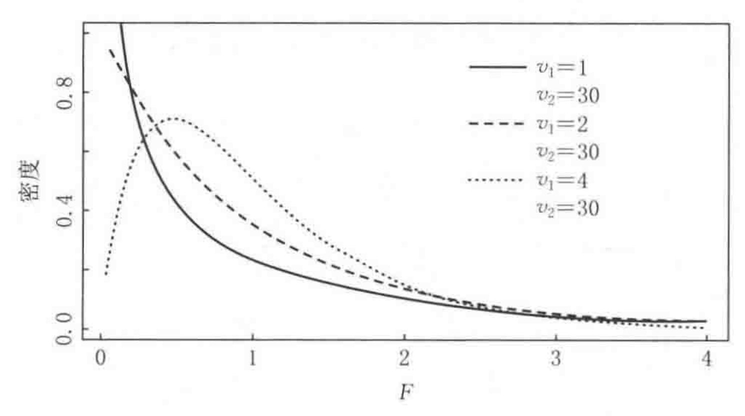


图 2.5 不同自由度(v_1 和 v_2)的 *F* 分布

第5节 | 从 F 到 t

当分子的自由度为1时，F 分布就等于 t 分布（其中 $v=v_2$）的平方。在第6章和第7章，我们会不断用到这一关系。这一事实串联起了我们前面回顾过的所有关系。首先，$t=\frac{Z}{\sqrt{\chi_v^2/v}}$，其中 $Z=\chi_1^2$，因此，当 $v_1=1$ 时，得到：

$$F=\frac{\chi_1^2/1}{\chi_{v_2}^2/v_2}=\frac{Z^2}{\chi_{v_2}^2/v_2}$$

所以，我们可以通过平方根把 F 转化为 t：

$$\sqrt{F}=\frac{Z}{\sqrt{\chi_v^2/v}}=t$$

第 6 节 | 小结

本章探讨了标准正态分布、χ^2 分布、t 分布和 F 分布之间的关系。由于检验统计量是随机分布的变量,所以理解它们的抽样分布对于我们理解如何运用功效分析至关重要。下一章将首先处理标准正态分布的功效分析,以便为其他分布的功效分析做好准备工作。

第3章

总体标准差已知的情况下，假设检验和功效分析的一般性议题：以两组均值为例

本章回顾基于总体的两组间均值差异的假设检验(z 检验)。为了讨论的简便,本章假定总体方差(σ^2)已知。在接下来的章节中,我们进一步探讨用样本(因此我们不得不对总体方差进行估计)来检验两组差异,这会影响分布的形状,进而影响功效。

第 1 节 | 总体标准差已知，均值差异服从正态分布的随机变量

推论统计之美在于，当总体标准差已知时，两组均值的差异也服从正态分布。其中，均值从随机样本中估计得到。已知有两个组，分别记为组 0（即控制组）和组 1（即干预组）。两组间的总体差异定义为 $\Delta=\mu_1-\mu_0$。其标准误（SE_Δ），即来自多个样本的组之间差异的分布之标准差，是这些样本所属总体的标准差（σ）的一个函数，包括总观测数（N）和每组所占比例（P_0 和 P_1，$P_0+P_1=1$，$1-P_0=P_1$）：

$$SE_\Delta=\frac{\sigma}{\sqrt{NP_1(1-P_1)}} \qquad [3.1]$$

上述表达式从概念上表明，抽样分布的标准差等于数据的标准差除以样本量相关函数的平方根。来自某个随机样本的变量 y 的组均值差异因此服从如下正态分布：

$$\bar{y}_1-\bar{y}_0\sim N(\Delta,\ SE_\Delta) \qquad [3.2]$$

第2节 | 总体标准差已知,两组均值差异的假设检验

假设检验的目的是,在给定总体假定的情况下,确定观察到特定数据的可能性。从性质上看,这个过程是先根据假定的零假设,把统计量抽样分布的形状中心化,然后计算我们手头的数据(或更为极端情况的数据)来自这一抽样分布的概率。如果观察到手头的数据(或更为极端情况的数据)的概率颇低,那么我们通常拒绝零假设,即零假设不大可能发生。其基本逻辑是,要么我们手头的数据错了,要么是零假设错了;然后我们基于零假设为真的假定,量化计算手头数据发生的可能性,据此作出抉择。

利用内曼—皮尔逊方法的假设检验

正如统计学家内曼和皮尔逊所言(Neyman and Pearson, 1933),假设检验是在两个假设之间作出抉择。零假设通常标记为 H_0,备择假设标记为 H_a。譬如,我们可以说零假设是组 0 和组 1 的均值相等。

$$H_0: \mu_1 = \mu_0$$

这是指两组均值的差异(Δ)为 0，即，

$$H_0: \mu_1 - \mu_0 = \Delta = 0$$

当我们检验组均值差异时，这就是典型的零假设。

考察两组均值时，备择假设(H_a)有三种可能形式。首先，可以提出备择假设为组 1 的均值大于组 0 的均值，

$$H_a: \mu_1 > \mu_0$$

即 $\Delta = \mu_1 - \mu_0$ 为正值，

$$H_a: \mu_1 - \mu_0 > 0$$
$$或\ H_a: \Delta > 0$$

其次，可以提出备择假设为组 1 的均值小于组 0 的均值，

$$H_a: \mu_1 < \mu_0$$

即 $\Delta = \mu_1 - \mu_0$ 为负值，

$$H_a: \mu_1 - \mu_0 < 0$$
$$或\ H_a: \Delta < 0$$

这两个备择假设就是所谓的单尾假设，下文会详述原因。最后，两组均值检验中常用的一个备择假设是组均值不相等，但不设定方向，

$$H_a: \mu_1 \neq \mu_0$$

即 $\Delta = \mu_1 - \mu_0$ 不等于 0，

$$H_a: \mu_1 - \mu_0 \neq 0$$
$$或\ H_a: \Delta \neq 0$$

这就是所谓的双尾检验，下文会详述为什么叫双尾检验。

假设检验所做的抉择在于:是否应拒绝零假设(H_0),而倾向于备择假设(H_a)? 拒绝零假设通常的办法是宣称,如果零假设为真,那么观测数据不可能发生,因为观察到的 Δ 与基于零假设的抽样分布所得到的值差异过大。

一种方法是假定 Δ 等于零假设所提出的值(一般来说,零假设宣称 $\Delta=0$),然后画出服从正态分布、均值为 $\Delta=0$、标准差为 SE_Δ 的抽样分布。切记,SE_Δ 是根据已知的总体标准差、样本量和每组样本量的比例计算得到的(方程 3.1)。换言之,我们可以利用样本量和已知的总体标准差,画出以零假设的假定值 Δ 为中心的正态曲线。

然后我们考虑,相比零假设,有多大可能观察到我们的 Δ。然而,这一方法不够标准化,而且取决于依据结果的测量单位画出抽样分布图的能力。虽然这一方法依然可行,但操作困难,而且难以很好地形成一般化的经验法则。

另一种可能方法是进行同样的流程,但不是基于 Δ 及其标准误来画抽样分布图,而是利用 Δ 及其标准误来计算标准化的检验统计量,然后利用标准正态分布,从零假设角度来考察手头数据发生的可能性。这一检验统计量是方程 2.2、方程 2.3 和方程 2.4 中所使用的 z 分数,即,

$$z=\frac{\Delta-\Delta_0}{SE_\Delta} \qquad [3.3]$$

其中 Δ_0 是零假设所假定的差异值,通常为 0。有了这一检验,我们就可以根据图 2.1 的标准正态分布,建立与第一类错误相应的拒绝域。

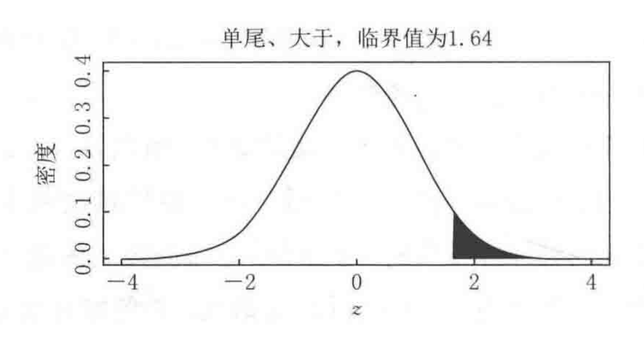
现在我们回到第 1 章所介绍的错误类别。首先是第一类错误,这是拒绝了零假设但零假设实际上为真的可能性。

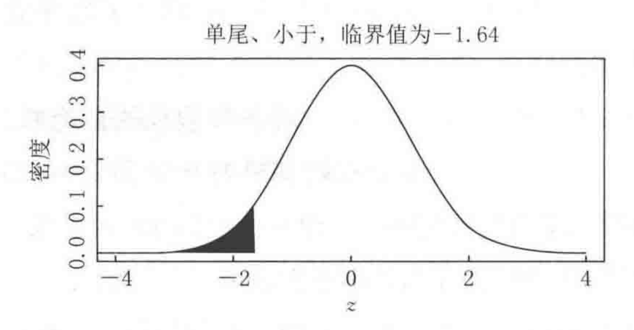
我们把第一类错误标记为 α，通常把 $\alpha=0.05$ 作为一个可接受的第一类错误发生可能性。

我们根据备择假设和第一类错误（α）取值，在基于零假设形成的抽样分布中建立拒绝域。这些拒绝域是基于零假设而建立的抽样分布中的一块区域，分布曲线下的这一区域面积所占比例代表了可接受的风险错误。拒绝域开始处的 z 值是临界值，如果检验统计量（方程 3.3）超过了临界值，那么就拒绝零假设。

图 3.1 呈现了 $\alpha=0.05$ 时三种备择假设的拒绝域。如你所见，前两幅图的阴影部分仅覆盖单侧 5%（即 $\alpha=0.05$）的区域面积。这就是它们被称为“单尾”检验的原因。第一幅图的阴影部分，即“大于”的备择假设 $\mu_1>\mu_0$，始于 $z=1.64$。这意味着如果方程 3.3 大于 1.64，我们就拒绝零假设，而倾向于接受备择假设。类似的逻辑也适用于图 3.1 中第二幅图的备择假设 $\mu_1<\mu_0$，为了拒绝零假设，检验统计量必须小于 -1.64。

最后，如果备择假设是均值不相等，即 $\mu_1\neq\mu_0$，我们就把 5% 的阴影区域分布到两侧，这也就是它们被称为“双尾”检验的原因。由于每一侧的阴影区域面积只占 2.5%，所以临界值就变得更大、更趋向于两端，即 $z=-1.96$ 和 $z=1.96$。如果检验统计量（方程 3.3）在任一方向的绝对值超过了 1.96，我们就拒绝零假设。当然，随着 α 取值的变化，比如 $\alpha=0.01$，抽样分布中拒绝域所占比例也会发生变化，因而临界值也会相应变化。在几乎所有的统计教科书中都可以找到不同 α 水平和备择假设所对应的临界值。





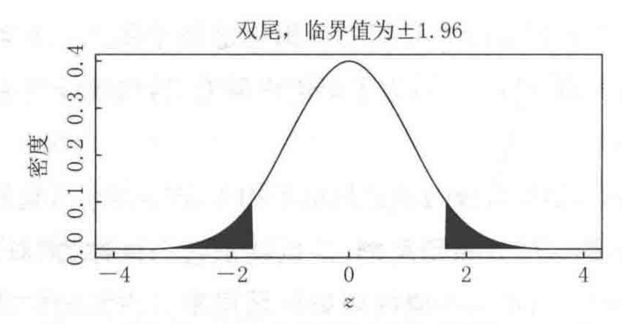


图 3.1　零假设 z 分布的单尾和双尾拒绝域($\alpha=0.05$)

什么是 p 值

多数统计软件程序不直接使用内曼—皮尔逊方法进行假设检验。这些软件在假定零假设为 $\Delta_0=0$ 的情况下，计算超出检验统计量取值的抽样分布两端区域的面积。譬如，若

检验统计量（方程3.3）为$z=3$，统计软件利用1减去方程2.4来计算数据检验结果（或更为极端的结果）在假定零假设为$\Delta=0$的情况下可能发生的概率：

$$p(z)=1-[\Phi(z)-\Phi(-z)]$$
$$p(z)=1-[\Phi(3)-\Phi(-3)]$$
$$p(z)=0.003$$

表示双尾p值的另一种方法是：

$$p(z)=2\times\left[1-\Phi\left(\frac{|\Delta|}{SE_\Delta}\right)\right] \qquad [3.4]$$

而单尾检验的p值表达式则为：

$$p(z)=1-\Phi\left(\frac{|\Delta|}{SE_\Delta}\right) \qquad [3.5]$$

使用p值进行检验背后的思想是，如果$p(z)$小于α，我们就认为差异“显著”。

尤为重要的是要认识到，p值并不是如果零假设为真，手头数据本身会发生的概率。数据只是连续分布上的一个点。连续分布上的概率需要一个区间，而单个点的概率则接近于0。因此，这也是为什么我们把p值说成是，如果零假设事实上为真，“数据检验结果或更极端结果”将会发生的可能性。

什么对功效分析有用

如下文所见，功效分析本质上是在给定不同区间取值情况下，计算曲线下的面积。因此，p值对功效分析没什么用，

因为我们永远不知道未来某数据所产生的精确 p 值。相反，我们可以提前规定 α 水平，用于计算临界值，帮助我们找到正态分布上特定区间的概率。如果总体标准差 σ 未知，需要我们对其进行估计，那么抽样分布就不再服从正态分布，而是服从其他分布了。

第 3 节 | 总体标准差已知的情况下，两组均值差异检验的功效分析

我们既然已经弄懂了如何利用正态分布来进行假设检验，那么接下来就可以进行功效分析了。功效分析与第二类错误(β)相关，即我们接受了零假设但实际上备择假设为真的可能性。思考第二类错误的一种方法是考虑我们的检验统计量小于 α 所规定临界值的可能性，类似于第 1 章的第二个模拟计算（见图 1.2）。例如，若备择假设为 $\Delta>0$，单尾检验设定 $\alpha=0.05$，则临界值 z 为 1.64。在收集数据之前，第二类错误想要问的是，即使备择假设为真，检验统计量（方程 3.3）小于 1.64 的可能性是多少？

为了回答这一问题，我们设想一些预期检验结果，然后围绕这些预期检验结果画出抽样分布。根据临界值的位置，单尾检验的第二类错误(β)就是小于右侧临界值的面积（若备择假设为正），或大于左侧临界值的面积（若备择假设为负）。双尾检验的第二类错误(β)是两个临界值之间的面积。而检验的功效就是这一区域之外的面积，即 $1-\beta$。因此，功效分析就是把基于备择假设为真而建立的抽样分布（备择分

布或非中心化分布),加到基于零假设为真而建立的抽样分布(虚无分布或中心化分布)之上。

为了计算备择分布上两个点在曲线下的区域面积,需要知道备择分布的均值和标准差。利用标准正态分布,我们可以把虚无分布曲线的标准差用于备择分布曲线(值为 1)。然而,由于虚无分布的均值为 0(因为零假设假定 $\Delta=0$),我们需要为备择分布找到一个合适的平均值。备择分布的均值就是预期的检验统计量,或当备择假设为真时方程 3.3 的期望值。通常把这一参数标记为 λ,称为非中心化参数(noncentrality parameter)。对方程 3.3 进行一些简单变换,运用 Δ 和方程 3.1,就得到了服从标准正态分布之检验统计量的非中心化参数的一般表达式:

$$\lambda=\frac{|\Delta|}{\sigma}\sqrt{NP_1(1-P_1)} \qquad [3.6]$$

需要注意的是,本书仅考察均值差异的绝对值。这使我们得以围绕正的临界值展开分析。

一旦有了 λ 的取值,我们就可以利用正态分布特性得到 $Z-\lambda \sim N(0, 1)$,因为服从标准正态分布的变量 Z 与非中心值 λ 的差值服从正态分布。利用这一特征的方法是运用标准正态分布的累积密度函数 Φ,来计算给定 α 水平下,小于临界值 $z_{1-\alpha}$ 的面积,用以估计单尾检验的第二类错误:

$$\beta=\Phi(z_{1-\alpha}-\lambda) \qquad [3.7]$$

或临界值之间的面积,用以估计双尾检验的第二类错误:

$$\beta=\Phi(z_{1-\alpha/2}-\lambda)-\Phi(z_{\alpha/2}-\lambda) \qquad [3.8]$$

一旦得到了第二类错误(β),则功效便是 $1-\beta$。

实例

假定某标准化检验中两组均值的差异期望值为 $\Delta=25$，且样本包含 100 个观测值，所在总体的标准差 $\sigma=75$。进一步假定这是一个均衡设计，即每组所含观测值皆为 50，因此 $P_1=50/100=0.5$。运用方程 3.6 可计算得到 λ：

$$\lambda=\frac{25}{75}\sqrt{100\times0.5\times(1-0.5)}=1.67$$

如果进行 $\alpha=0.01$ 的单尾检验，则标准正态分布的临界值为 2.326；或者说当 $1-\alpha=1-0.01=0.99$ 时，对应 z 分布上的取值 $z_{1-\alpha}=z_{0.99}=2.326$。然后就可以计算第二类错误，先计算 $z_{1-\alpha}-\lambda=2.326-1.667=0.659$，再求出标准正态分布上小于标准值 0.659 的面积，即 $\beta=\Phi(0.659)=0.745$。这意味着功效约为 $1-\beta=1-0.745=0.225$。这一单尾检验如图 3.2 所示，其中的灰色阴影面积约占备择分布的 75%（即 $\beta\times100$），备择分布中右侧临界值（即黑色阴影部分起始处）往右的面积就是检验的功效。

如果使用相同的 λ 值进行 $\alpha=0.01$ 的双尾检验，则临界值 z 为 2.576；或者说当 $1-\alpha/2=1-0.01/2=0.995$，对应 z 分布上的取值 $z_{1-\alpha/2}=z_{0.995}=2.576$。第二类错误就等于正态分布上 $z_{1-\alpha/2}-\lambda$ 和 $z_{\alpha/2}-\lambda$ 两个值之间的面积，即 $\beta=\Phi(2.576-1.667)-\Phi(-2.576-1.667)=0.818$。这意味着功效约为 $1-\beta=1-0.818=0.182$。该双尾检验如图 3.3 所示，其中灰色阴影面积约占备择分布的 82%（即 $\beta\times100$），而左右两侧临界值（即黑色阴影部分起始处）以外的面积就是

检验的功效。

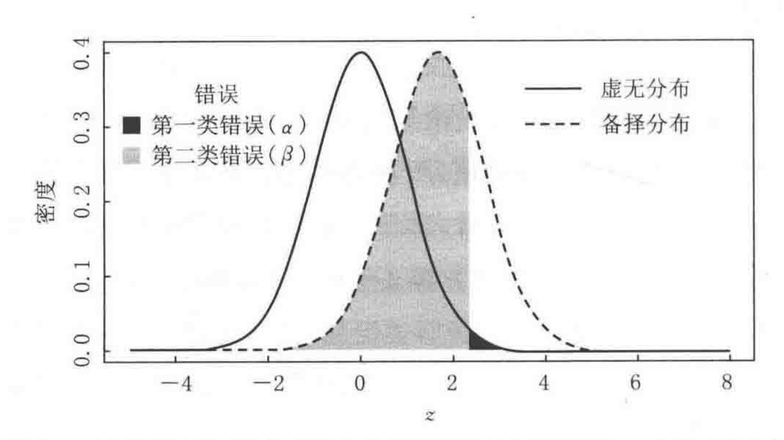


图 3.2 单尾检验的服从标准正态分布的虚无分布(实线)($\alpha=0.01$,临界值 $z_{\alpha/2}=2.326$)以及非中心化参数 $\lambda=1.667$ 和 $\beta=0.745$ 的备择分布(虚线)

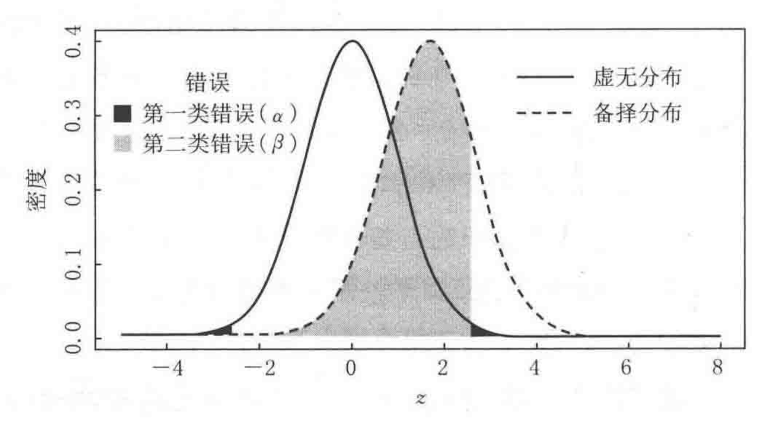


图 3.3 双尾检验的服从标准正态分布的虚无分布(实线)($\alpha=0.01$,临界值 $z_{\alpha/2}=\pm2.576$)以及非中心化参数 $\lambda=1.667$ 和 $\beta=0.818$ 的备择分布(虚线)

有用的算式关系

基于标准正态分布,我们利用方程 3.7 和方程 3.8 可以得到一些有用的算式关系。譬如,把单尾检验的第二类错误

记为 $\beta=\Phi(z_{1-\alpha}-\lambda)$。其中 β 是正态曲线下的一块面积。因此，我们就可以在等式的两边都消掉 Φ，得到 $z_\beta=z_{1-\alpha}-\lambda$，进行简单变换之后就得到 $\lambda=z_{1-\alpha}-z_\beta$。这表明在标准正态检验情况下，检验统计量的期望值是标准正态分布上两个分位数（分别在 $1-\alpha$ 和 β 上）的函数。双尾检验的情况更为复杂一些，因为增加了两个累积密度函数来计算 β。但第二项 $\Phi(z_{\alpha/2}-\lambda)$ 的值通常很小*，因为这是检验分布中左侧临界值再往左的面积。在上面的实例中，这一项的值为 0.000 01。如果这一项忽略不计，我们就可使用与单尾检验相同的方法得到双尾检验的 $\lambda\approx z_{1-\alpha/2}-z_\beta$。因此，无论是单尾检验还是双尾检验，我们都可以把临界值的绝对值标记为 $z_{critical}$，从而得到一般化的算式关系：

$$\lambda\approx z_{critical}-z_\beta \qquad [3.9]$$

如果 $\alpha=0.05$，则双尾检验的 $z_{critical}=1.96$，单尾检验的 $z_{critical}=1.68$。若功效为 0.8，则 $z_\beta=z_{0.2}=-0.84$；若功效为 0.9，则 $z_\beta=z_{0.1}=-1.28$。譬如，功效为 0.8 的 α 双尾检验的检验统计量期望值约为 $1.96-(-0.84)=2.8$，如下文所示，在不同类型的功效分析中经常会用到这一数字。

分布分位数

$z_{critical}$ 和 z_β 是标准正态分布上的分位数取值。分位数是累积密度函数的倒数，是给定累积概率的 z 的取值。可用计算机程序来计算正态分布的分位数。为了使用便利，本书附

---

* 即双尾检验 $\beta=\Phi(z_{1-\alpha/2}-\lambda)-\Phi(z_{\alpha/2}-\lambda)$ 的计算式中的第二项。——译者注

录表 A.4 的最后一行收录了标准正态分布的一些重要分位数。

附录表 A.4 的左边几列在计算与 β 值相关的正态分布分位数时非常有用。譬如,第二类错误为 0.1 时,则使用 $z_{\beta}=z_{0.1}=-1.282$。而附录表 A.4 的右边几列则可用于计算与检验类型相关的临界值绝对值。譬如,若 $\alpha=0.05$ 的单尾检验,则使用 $z_{1-\alpha}=z_{0.95}=1.645$;若为双尾检验,则使用 $z_{1-\alpha/2}=z_{0.975}=1.96$。

第 4 节 | 无标尺参数

行文至此，功效分析几乎所有的组成要素都是以结果变量的量纲为测量单位的。μ_1、μ_0、Δ 和 σ 的值都是以所讨论变量为单位，无论该变量是标准化测验、体质指数还是"小蜜蜂"分数*。两组均值差异的实际大小很难预测，而样本中的方差同样难以预测。如果我们处理的是所谓的"无标尺"参数，那么功效分析就会更为简单。本书介绍的第一个无标尺参数是效应值(effect size)。科恩(Cohen，1988)的 d 就是一个效应值，它通过除以总体标准差 σ 对组均值差异 Δ 进行标准化。这一度量可用小写的希腊字母 δ 表示，其计算式如下所示：

$$\delta=\frac{\mu_1-\mu_0}{\sigma}=\frac{\Delta}{\sigma} \qquad [3.10]$$

也就是方程 3.6 乘式中的第一项。

δ 的含义相对比较清晰，即以标准差单位度量的均值差异。因此，$\delta=0.5$ 表示，平均而言，其中一组要比另一组大半个标准差。用 δ 替代 $\frac{\Delta}{\sigma}$ 的原因有二。一是使研究者以更为

* "小蜜蜂"(Galaga)，一种固定射击的街机游戏。——译者注

一般化的方式来思考未来研究的结果,即以标准单位度量,组间的一般化差异是多少?

第二个原因可能更为重要,即效应值允许研究者利用已有研究进行预估(参见第8章),即使这些已有研究使用了不同的度量。譬如,假定研究者正在计划研究为上大学建立的儿童储蓄账户如何影响父母的理财素养。一项已有研究可能使用了量表A,而当下的研究者想要使用量表B。如果量表A的分数区间为1—100,而量表B的取值区间为0—7,那就很难把量表A得到的组间差异转化到量表B。然而,如果这项已有研究公布了干预组和控制组的均值和标准差,那就可以确定一个效应值(下文会详述)。有了效应值,使用量表B的研究者就能对其研究中的预期效应有所预判。

所以,可以用效应值来表示非中心化参数,如下式:

$$\lambda = \underbrace{\delta}_{\text{效应值}} \underbrace{\sqrt{NP_1(1-P_1)}}_{\text{样本量}} \qquad [3.11]$$

在未来的研究中,即使没有标尺化参数的相关信息,利用上式也可以进行功效分析。

第 5 节 | 均衡还是非均衡？

研究设计中经常遇到的另一个议题是均衡问题。均衡研究中每组都有相同的个案数，而非均衡研究中的每个组所包含的个案数各不相同。均衡曾经是研究设计中的一个重要考量，尤其是在现代计算出现以前，因为均衡设计可以简化计算。如今，研究者最终处理的数据几乎不可能是完全均衡的，但或许接近均衡。

但在研究设计中，均衡依然是一个重要的考量，原因如下。首先，总样本量固定的情况下，均衡样本的功效要大于非均衡样本。其次，如下文将要阐述的，均衡样本中的非中心化参数及其计算更为简便。最后，有时均衡设计可能不合人意，因为我们希望控制组较小，让研究对象中尽可能多的个案进入干预组。

非均衡设计中的非中心化参数

我们已经讨论了非均衡设计情况下的非中心化参数（方程 3.11）。在这种情况下，组 1 的样本量为 n_1，组 0 的样本量为 n_0，总样本量为 $n_1+n_0=N$，则组 1 个案数所占比例为 $n_1/N=P_1$。因而，非均衡设计中的计算需要用到效应值

δ、总样本量 N 和其中一组个案数所占的比例(我更偏好组 1 和 P_1,不过选择哪一组都一样)。

均衡设计中的非中心化参数

当 $n_1=n_0=n$ 时,非中心化参数的计算更为简化。此时,$P_1=(1-P_1)=0.5$,则非中心化参数(方程 3.11)中的第二项为 $\sqrt{N\times0.5^2}=\sqrt{N\times0.25}=\sqrt{N/4}=\sqrt{n/2}$。因而均衡设计中非中心化参数的计算就变为:

$$\lambda=\underbrace{\delta}_{\text{效应值}}\underbrace{\sqrt{n/2}}_{\text{样本量}} \qquad [3.12]$$

所以均衡设计中的计算仅需要效应值 δ 和每组样本量 n 的信息。

第6节 | 功效分析的类型

在进行某研究的样本设计时，通常会提出三类问题：

(1) 计算先验功效(power a priori)：给定 α、效应值和样本量的情况下，功效是多少？

(2) 计算最低样本量：给定功效水平 $(1-\beta)$、α 和效应值的情况下，需要多大的样本量？

(3) 计算可检测的最低效应值：给定功效水平 $(1-\beta)$、α 和样本量的情况下，可检测的最低效应值是多少？

计算先验功效

本书至今为止主要是在讨论给定效应值、样本量和 α 水平的情况下计算功效。简要回顾一下，当已知如方程 3.10 所定义的期望效应值 δ、总的样本量 N、某组的样本比例 P 和 α，就能运用标准正态分布的累积密度函数得到某单尾检验的第二类错误(β)：

$$\beta=\Phi(z_{1-\alpha}-\lambda)=\Phi[z_{1-\alpha}-\delta\sqrt{NP_1(1-P_1)}]$$

双尾检验的计算过程也类似：

$$\beta=\Phi(z_{1-\alpha/2}-\lambda)-\Phi(z_{\alpha/2}-\lambda)$$

$$\beta=\Phi[z_{1-\alpha/2}-\delta\sqrt{NP_1(1-P_1)}]-\Phi[z_{\alpha/2}-\delta\sqrt{NP_1(1-P_1)}]$$

其中,z_a 是标准正态分布在点 a 上的分位数。

当数据被认为均衡时,上述计算式可以进行如下简化。就单尾检验而言,

$$\beta=\Phi(z_{1-\alpha}-\lambda)=\Phi(z_{1-\alpha}-\delta\sqrt{n/2})$$

对双尾检验而言,

$$\beta=\Phi(z_{1-\alpha/2}-\lambda)-\Phi(z_{\alpha/2}-\lambda)$$

$$\beta=\Phi(z_{1-\alpha/2}-\delta\sqrt{n/2})-\Phi(z_{\alpha/2}-\delta\sqrt{n/2})$$

这些计算要求拥有计算标准正态分布的累积密度函数功能的计算机或临界值表。无论如何,只要确定了 β 值,该检验的功效便为 $1-\beta$。

计算最低样本量

为了计算最低样本量,我们回到前文论及之有用的算式关系,即 $\lambda\approx z_{critical}-z_\beta$。利用这一关系对非中心化参数 λ 进行拆解,就能计算得到特定 α 水平下要检测到特定效应值(得到 $z_{critical}$)和功效(得到 z_β)的最低样本量。

非均衡设计

在非均衡设计中,总体标准差已知的情况下,检验两组差异的非中心化参数(方程 3.11)是 $\lambda=\delta\sqrt{NP_1(1-P_1)}$,这一计算式可以写成如下表达式:

$$\delta\sqrt{NP_1(1-P_1)}\approx z_{critical}-z_\beta$$

对上式进行变换可求得 N，得到如下表达式：

$$N \approx \frac{(z_{critical} - z_{\beta})^2}{[P_1(1-P_1)]\delta^2} \qquad [3.13]$$

在这一表达式中，若功效和显著性水平（分子）保持不变，样本量会随着 P_1 值远离 50% 而增加，随着效应值（δ）增加而减小。

实例

譬如，要进行一项双尾检验，其中 $\alpha = 0.05$，功效为 0.8，效应值 $\delta = 0.3$，组 1 样本量占总样本量的 1/4，因此得到 $z_{critical} = z_{1-\alpha/2} = z_{0.975} = 1.96$，$\beta = 1 - 0.8 = 0.2$，$z_{\beta} = z_{0.2} = -0.842$，$P_1 = 0.25$。把这些值代入方程 3.13，得到：

$$N \approx \frac{(z_{critical} - z_{\beta})^2}{[P_1(1-P_1)]\delta^2} \approx \frac{[1.96-(-0.842)]^2}{[0.25(1-0.25)]0.3^2} \approx 465.26$$

即所需总样本量约为 466（取整时一般向上取较大的整数）。

均衡设计

如果观测样本是均衡的，则 $n_1 = n_0 = n$，$2 \times n = N$，非中心化参数也简化为方程 3.12，得到 $\lambda = \delta\sqrt{n/2} \approx z_{critical} - z_{\beta}$，进行变换之后得到：

$$n \approx \frac{2(z_{critical} - z_{\beta})^2}{\delta^2} \qquad [3.14]$$

就像非均衡设计一样，这一表达式说明样本量会随着分母中效应值（δ）的增加而减小。

实例

譬如,要进行一项双尾检验,其中 $\alpha=0.05$,功效为 0.8,效应值 $\delta=0.3$,因此得到 $z_{critical}=z_{1-\alpha/2}=z_{0.975}=1.96$, $\beta=1-0.8=0.2$, $z_{\beta}=z_{0.2}=-0.842$。把这些值代入方程 3.14,得到:

$$n\approx\frac{2(z_{critical}-z_{\beta})^2}{\delta^2}\approx\frac{2[1.96-(-0.842)]^2}{0.3^2}\approx174.47$$

即每组个案数约为 175,总样本量 $2n=N=350$。值得注意的是,本例中的所有参数取值都与非均衡设计中一样,但所需的总样本量 N 更小。

计算最低可检测效应值

最低可检测效应值(Bloom, 1995)是一个建构的度量,用以概括某个样本的灵敏度。这是在假定样本量和 α 水平不变的情况下,在特定功效水平上能检测到的最小效应值(δ)。在此回到前文所述有用的算式关系部分,即 $\lambda\approx z_{critical}-z_{\beta}$。利用这一关系和非中心化参数 λ 的分解,就能计算出给定样本量、α 水平(计算 $z_{critical}$)和功效(计算 z_{β})的情况下得到显著检验结果的效应值。

非均衡设计

对非中心化参数(方程 3.11)和($z_{critical}-z_{\beta}$)进行变换,就能得到最低可检测效应值:

$$\delta_m \approx \frac{z_{critical} - z_\beta}{\sqrt{NP_1(1-P_1)}} \quad [3.15]$$

这一表达式说明可检测效应会随着样本量 N 的增大而减小，但会随样本变得更不均衡而增加（即 P_1 远离 0.5）。

实例

譬如，要进行一项双尾检验，其中 $\alpha=0.05$，功效为 0.8，总样本量 $N=500$，其中组 1 样本量占总样本量的 1/4。得到 $z_{critical}=z_{1-\alpha/2}=z_{0.975}=1.96$，$\beta=1-0.8=0.2$，$z_\beta=z_{0.2}=-0.842$，$P_1=0.25$。把这些值代入方程 3.15，得到：

$$\delta_m \approx \frac{z_{critical} - z_\beta}{\sqrt{NP_1(1-P_1)}} \approx \frac{1.96-(-0.842)}{\sqrt{500\times0.25(1-0.25)}} \approx 0.289$$

这一效应值比上一个例子中的 0.3 略小，因为样本量从 466 增加到了 500。

均衡设计

如果观测样本是均衡的，则 $n_1=n_0=n$，$2\times n=N$，非中心化参数可变换为：

$$\delta_m \approx (z_{critical} - z_\beta)\sqrt{\frac{2}{n}} \quad [3.16]$$

这一表达式说明，可检测效应会随着样本量的增大而减小。

实例

譬如，要进行一项双尾检验，其中 $\alpha=0.05$，功效为 0.8，每个组的样本量 $n=175$，求效应值。得到 $z_{critical}=z_{1-\alpha/2}=$

$z_{0.975}=1.96$, $\beta=1-0.8=0.2$, $z_{\beta}=z_{0.2}=-0.842$。* 把这些值代入方程 3.16,得到:

$$\delta_m \approx (z_{critical}-z_{\beta})\sqrt{\frac{2}{n}} \approx [1.96-(-0.842)]\sqrt{\frac{2}{175}} \approx 0.300$$

即当每个组的样本量为 175 时,效应值为 0.3。

* 原书有误,译者进行了部分修改。——译者注

第 7 节 | 功效表

本书关注的是如何使用公式来计算结果，这就需要我们了解软件的使用。在功效分析软件广泛使用之前，多数研究者依靠提供功效表格的书籍，包括科恩(Cohen，1988)的开创性著作。这些表格使研究者可以进行三种类型的功效分析(当然，需要花些工夫)。表 3.1 就是这样一个例子，该表源自科恩(Cohen，1988)著作中一张类似的表格。

表 3.1　$\alpha=0.05$ 水平下双尾检验的功效表

n	δ				
	0.1	0.2	0.3	0.4	0.5
50	8	17	32	51	70
52	8	17	33	52	71
54	8	18	34	54	73
56	8	18	35	55	75
58	8	19	36	57	76

注：本表与本书作者使用软件计算得到的功效值之间的微小差异源于科恩(Cohen，1988)使用了近似值。

资料来源：Cohen，1988，p.37，Table 2.3.5。

功效表主要围绕三类关键信息而构建：样本量、效应值和功效。不同的表格则根据其他的假定制作，如单尾还是双尾、第一类错误(α)。利用表 3.1，读者可以根据行来选择样本量，根据列来选择效应值，然后在特定的格子中得到功效。

如表 3.1 所示,对于每组样本量为 58(最后一行)、效应值为 0.4(倒数第二列)的某个研究,其 $\alpha=0.05$ 水平下双尾检验的功效为 0.57(注意,功效表一般不显示小数点)。

第8节 | 小结

从理解抽样分布的视角出发,本章聚焦于如何进行功效分析,并区分了两种情况:一是零假设为真(中心化分布),二是备择假设为真(非中心化分布或备择分布)。如果总体标准差已知,可使用标准正态分布及其分位数来进行功效分析。我也介绍了效应值以及对无标尺参数的需求。下一章中,我们将讨论更为常见的情形,即总体标准差未知,必须通过估计得到。这意味着必须利用 t 分布(Student, 1908),该分布取决于自由度,而自由度又取决于样本量。由于样本量对于功效分析如此重要,我们就会发现它使分析过程变得复杂。

第4章

总体标准差未知、需要估计的情况下，来自简单随机样本的两组差异

本章讨论组间均值差异的检验,但现在我们认识到总体标准差未知,必须通过估计得到。这改变了检验和确定功效所必须使用的分布。如今必须使用 t 分布而非标准正态分布(Student, 1908)。由于 t 分布取决于样本量,所以功效分析就没那么简单直观了。本章将聚焦于组间均值差异检验的普通最小二乘回归方法。这为后面增加协变量和聚类设计奠定了基础。

我们使用一个研究早餐与增重关系实验的数据,作为均衡设计的一个操作实例。该研究名为"早餐建议对减肥的有效性",由杜兰达及其同事(Dhurandhar et al., 2014)牵头执行,是一个随机控制试验,用于评估不同的早餐进食方式如何影响减肥。该研究对一个研究样本随机指定干预条件,研究样本是一群肥胖但其他方面都健康的个体。[4] 干预条件包括两组,其中一组被告知要吃早餐,另一组被告知不要吃早餐;而控制组仅给予营养指南。该研究发现,虽然干预组的早餐习惯存在差异,但两者在身体质量指数(简称体质指数,BMI)上的差异统计不显著。体质指数被定义为体重(千克)除以身高(米)的平方,单位为千克/米^2。

基于本书的主旨,我们仅比较吃早餐和不吃早餐两组

个案。* 我们使用研究参与者在研究期结束前测得的 BMI 值作为结果变量。该研究在 clinicaltrials.gov 的注册编号是 NCT01781780，相关文件包括了功效分析。

表 4.1　干预后体质指数的概要统计

	N	均值	标准差
不吃早餐(控制组)	25.000	35.340	5.966
吃早餐(干预组)	25.000	31.459	5.476
总样本	50.000	33.400	5.997

表 4.1 呈现了我们所需的子样本重要信息(均值、标准差和样本量)。分析所使用的原始数据参见附录。虽然本书仅使用了实际数据的很小一部分，但所有的公开数据可在 ICPSR[5] 获取(研究编号为 36174)。

---

* 读者需要注意的是，如表 4.1 所示，出于两组比较的目的，本书作者把不吃早餐作为控制组，吃早餐作为干预组，这一操作与“早餐建议对减肥的有效性”研究所设置的干预组和控制组有所不同。——译者注

第 1 节 | 数据产生过程

我们首先介绍实验学者所谓的实验设计模型方程(Kirk, 1995)。其他研究领域认为这就是“数据产生过程”。我们假定的模型是:

$$y_{ij}=\mu+\tau_j+e_{ij} \qquad [4.1]$$

其中 y_{ij} 是结果变量在实验组 j 中的第 i 个观测值,该实验组来自方差为 σ^2 的总体,μ 是结果变量在所有观测值上的均值,τ_j 是组 j 的干预效应[6](其限制是 $\sum_j \tau_j=0$),而 e_{ij} 则是与实验组 j 中的第 i 个观测值相关的误差。每个实验组都包含 $i=\{1, 2, 3, \cdots, n_j\}$ 个观测值,而实验组的数量为 $j=\{1, 2, \cdots, p\}$。本书中,我们只关心两个组的情况,即 $p=2$。此外,我们把干预组标记为 $j=1$,控制组为 $j=0$,从而与干预指示变量的编码保持一致。因此,控制组 $j=0$ 的样本量为 n_0,而干预组 $j=1$ 的样本量为 n_1,总样本量为 $N=n_0+n_1$。*

---

* 本书作者有时用“treatment group”同时指代受到干预的组和控制组,有时仅指受到干预的组别。因此,为了表达的清晰和准确,我根据语境分别把上述两种情况译为实验组和干预组,实验组包含干预组和控制组两种情况,而干预组仅指受到干预的组别。——译者注

第2节 | 检验样本组间均值差异

像组间均值差异这类随机变量的抽样分布由均值和方差决定。这一部分将以回归视角来考察均值差异以及均值差异的抽样方差。我们分别考虑非均衡设计和均衡设计两种情况，前者指每个组的观测数不同（即 $n_0 \neq n_1$），后者指每组观测数相同（即 $n_0 = n_1$）。

结果变量的总均值 μ 由各组所有观测值总和除以总样本量估计得到：

$$\bar{y} = \frac{\sum_j \sum_i y_{ij}}{N} \qquad [4.2]$$

组 j 的均值估计值 μ_j 由下式估计得到：

$$\bar{y}_j = \frac{\sum_i^{n_j} y_{ij}}{n_j} \qquad [4.3]$$

干预效应 τ_j 等于组均值与总均值的差异，即 $\tau_j = \mu_j - \mu$。由于假定总体中只有 p 种干预可供选择，因此我们把干预效应看作是对干预总体的固定选择，而非随机选择。如果处理的是对可能干预的随机选择，分析会有所不同。最后，e_{ij} 是组内残差，即组 j 的第 i 个观测值与组 j 的均值之差，$e_{ij} = y_{ij} - \mu_j$。由于干预效应是唯一的协变量，残差的方差就是未引入

干预时总体的方差,故而 e_{ij} 的方差就是 σ^2。

正如下文将看到的,与第 3 章类似,均值差异的抽样方差之分布的方差如方程 4.15 所示,取决于对总体方差 σ^2 的估计,而总体方差是未知的,必须从数据中估计得到。调查数据中经常用到的一种方差估计方法是:

$$\tilde{\sigma}^2 = \frac{\sum_j \sum_i (y_{ij} - \bar{y})^2}{N-1} \qquad [4.4]$$

但这不是总体方差的一个优良估计,因为引入干预就等于在数据中引入了变异(我们也希望如此)。一个更好的估计方法是用组均值 $\bar{y}_j$ 替代总均值 $\bar{y}$,并修正分母中的自由度:

$$\hat{\sigma}^2 = \frac{\sum_j \sum_i (y_{ij} - \bar{y}_j)^2}{N-p} \qquad [4.5]$$

这一方程得到的便是概论性统计书籍中所说的合并方差,一般定义为每组中因变量值的方差的加权均值。在两个组别的情况下,其表达式为:

$$\hat{\sigma}^2 = \frac{(n_1-1)\hat{\sigma}_1^2 + (n_0-1)\hat{\sigma}_0^2}{n_1+n_0-2} \qquad [4.6]$$

其中干预组的方差被定义为:

$$\hat{\sigma}_1^2 = \frac{\sum_i (y_{i1} - \bar{y}_1)^2}{n_1 - 1}$$

控制组的方差被定义为:

$$\hat{\sigma}_0^2 = \frac{\sum_i (y_{i0} - \bar{y}_0)^2}{n_0 - 1}$$

回归模型

简单一元回归使用最小二乘法标准来拟合一条最契合数据的直线，以描述两个变量之间的关系。可以把回归分析模型看作[7]：

$$y_{ij}=\gamma_0+\gamma_1 T_{ij}+e_{ij} \qquad [4.7]$$

其中 y 是感兴趣的线性结果变量，T 是干预解释变量。[8]

为了估计回归斜率 γ_1，多数概论性统计书都会提供下述等式：

$$\hat{\gamma}_1=\frac{\sum_j \sum_i (y_{ij}-\bar{y})(T_{ij}-\bar{T})}{\sum_j \sum_i (T_{ij}-\bar{T})^2} \qquad [4.8]$$

其中$\bar{T}$是干预指示变量的均值。* 在两组的情况下，干预指示变量的通常编码方式是二分类或“虚拟”变量：

$$T_{ij}=\begin{cases}0,\text{如果属于控制组}\\1,\text{如果属于干预组}\end{cases} \qquad [4.9]$$

根据这一编码，T 的均值就很简单：

$$\bar{T}=\frac{\sum_j \sum_i T_{ij}}{N}=\frac{n_1}{N}=P_1 \qquad [4.10]$$

当 T 是二分类指示变量时，将会看到斜率 γ_1 是所有 $T=1$ 个案均值与所有 $T=0$ 个案均值之差的估计值，$\gamma_1=\bar{y}_1-\bar{y}_0$。

* 在两组比较的情况下，干预指示变量（或干预变量）即识别干预组和控制组的二分变量。——译者注

一元回归模型的截距之估计也非常直观:

$$\hat{\gamma}_0 = \bar{y} - \hat{\gamma}_1 \bar{T} \qquad [4.11]$$

后面我们会看到这一式子的值就等于组 0 的均值。

在一元回归中,斜率 γ_1 的估计抽样方差通常表示为:

$$\hat{\text{var}}\{\hat{\gamma}_1\} = \frac{\hat{\sigma}_e^2}{\sum_j \sum_i (T_{ij} - \bar{T})^2} \qquad [4.12]$$

其中,一元回归的均方误差(MSE 或 σ_e^2)之估计方法为:

$$\hat{\sigma}_e^2 = \frac{\sum_j \sum_i \hat{e}_{ij}^2}{N-2} \qquad [4.13]$$

在回归分析框架中,e 是观测取值 y_{ij} 与其拟合值$\hat{y}_{ij}$之间的差,$e_{ij} = y_{ij} - \hat{y}_{ij}$。在两组的情况下,其实只有两个拟合值:

$$\hat{y}_{ij} = \begin{cases} \bar{y}_0, \text{如果 } T_{ij} = 0 \\ \bar{y}_1, \text{如果 } T_{ij} = 1 \end{cases} \qquad [4.14]$$

这意味着,在两组情况下,均方误差等于*:

$$\hat{\sigma}_e^2 = \frac{\sum_i^{n_0} (y_{i0} - \bar{y}_0)^2 + \sum_i^{n_1} (y_{i1} - \bar{y}_1)^2}{n_0 + n_1 - 2}$$

求均值差异的抽样方差(方程 4.12)的分子的另一种方法是使用 y 的分组方差 $\hat{\sigma}_j^2 = \frac{\sum_i^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1}$,从而得到 $\sum_i (y_{ij} - \bar{y}_j)^2$ 等于$(n-1)\hat{\sigma}_j^2$。因此得到均方误差为:

$$\hat{\sigma}_e^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_0 - 1)\hat{\sigma}_0^2}{n_0 + n_1 - 2}$$

---

* 原文中下式的下标有误,已修改。——译者注

这说明对于预测变量为二分变量的一元回归模型而言，均方误差的估计值等于总体方差 σ^2 的估计值。

抽样方差（方程 4.12）的分母 $\sum_j \sum_i (T_{ij} - \bar{T})^2$ 为：

$$\sum_j \sum_i (T_{ij} - \bar{T})^2 = n_0 \bar{T}^2 + n_1(1 - 2\bar{T} + \bar{T}^2)$$
$$= N\bar{T}(1 - \bar{T})$$

因此，γ_1 的抽样方差的估计值为：

$$\hat{\text{var}}\{\hat{\gamma}_1\} = \frac{\hat{\sigma}^2}{N\bar{T}(1-\bar{T})} \qquad [4.15]$$

其平方根即标准误：

$$SE_{\gamma_1} = \frac{\hat{\sigma}}{\sqrt{N\bar{T}(1-\bar{T})}}$$

由于 $\bar{T} = P_1$，所以上式就等同于方程 3.1。

非均衡设计的斜率与截距

利用代数计算，可以发现非均衡设计中 T 的斜率就是组均值差异。运用方程 4.10，得到：

$$\hat{\gamma}_1 = \frac{\sum_i^{n_1} y_{i1} - \sum_i^{n_1} y_{i1}\bar{T} - n_1\bar{y} + n_1\bar{T}\bar{y} + \bar{T}(n_0\bar{y} - \sum_i^{n_0} y_{i0})}{n_1 - 2n_1\bar{T} + n_0\bar{T}^2 + n_1\bar{T}^2}$$

$$\hat{\gamma}_1 = \frac{\sum_i^{n_1} y_{i1}}{N\bar{T}} - \frac{\sum_i^{n_0} y_{i0}}{N(1-\bar{T})} = \frac{\sum_i^{n_1} y_{i1}}{n_1} - \frac{\sum_i^{n_0} y_{i0}}{n_0} = \bar{y}_1 - \bar{y}_0$$

模型截距（方程 4.11）是控制组 0 的均值，理解了 $\gamma_1 = \bar{y}_1 - \bar{y}_0$，且 $\bar{y} = \bar{T}\bar{y}_1 + (1-\bar{T})\bar{y}_0$，我们就可以得到：

$$\gamma_0=\bar{y}-\bar{T}\gamma_1=\bar{T}\bar{y}_1+(1-\bar{T})\bar{y}_0-\bar{T}(\bar{y}_1-\bar{y}_0)=\bar{y}_0$$

均衡设计的斜率与截距

均衡设计中斜率和截距的含义更为直观。当认识到每组包含个案数相同,因而 T 的均值 $\bar{T}$ 为 0.5,$(T_{ij}-\bar{T})^2$ 为 0.25 时,我们就能理解斜率的含义了。这意味着方程 4.8 的分母是总样本量除以 4,或半数样本量 n 除以 2,即,

$$\sum_j \sum_i (T_{ij}-\bar{T})^2=\frac{N}{4}=\frac{n}{2} \qquad [4.16]$$

而方程 4.8 的分子则简化为 $\frac{1}{2}\sum_i(y_{i1}-\bar{y})-\frac{1}{2}\sum_i(y_{i0}-\bar{y})$。因此斜率可以表示为:

$$\hat{\gamma}_1=\frac{\frac{1}{2}\sum_i(y_{i1}-\bar{y})-\frac{1}{2}\sum_i(y_{i0}-\bar{y})}{n/2}$$

$$=\frac{\sum_i(y_{i1}-\bar{y})-\sum_i(y_{i0}-\bar{y})}{n}$$

进一步计算得到:

$$\hat{\gamma}_1=\frac{\sum_i y_{i1}}{n}-\frac{\sum_i y_{i0}}{n}=\bar{y}_1-\bar{y}_0$$

关于截距,假定的虚拟编码 T 的均值为 0.5,且由于总均值 $\bar{y}$ 是组均值的平均值,因而就可以直观地发现,截距是 $T=0$ 这一组的均值:

$$\hat{\gamma}_0=\frac{\bar{y}_1+\bar{y}_0}{2}-\frac{\bar{y}_1-\bar{y}_0}{2}=\bar{y}_0$$

t 检验

需要估计未知参数 σ 的分析必须使用 t 分布，而不能使用正态分布及其累积分布函数。在检验中使用 t 分布的可能影响是，假设检验中临界值的绝对值会更大，且随样本量变化。譬如，若样本量 $N=20$，则自由度 $v=df=20-2=18$。利用计算机程序或表格就能求出临界值，对于 $\alpha=0.05$ 且 $v=df=18$ 情况下的双尾检验，t 的临界值为 2.10，这要大于使用正态分布时的临界值 1.96（参见第 2 章）。

然而，t 检验的计算过程与 z 检验基本相同。均值差异也相同。均值差异的估计抽样方差是（方程 4.15）：

$$\hat{\text{var}}\{\bar{y}_1-\bar{y}_0\}=\frac{\hat{\sigma}^2}{N\bar{T}(1-\bar{T})} \qquad [4.17]$$

其平方根即标准误：

$$SE_{\bar{y}_1-\bar{y}_0}=\frac{\hat{\sigma}}{\sqrt{N\bar{T}(1-\bar{T})}}$$

除了现在使用的是 σ 的估计值之外，该表达式与方程 3.1 相同。如果两组样本量相同，上述表达式就可以进一步简化为仅用样本量表示：

$$\hat{\text{var}}\{\bar{y}_1-\bar{y}_0\}=\hat{\sigma}^2\frac{2}{n} \qquad [4.18]$$

抽样方差的平方根就是均值差异的标准误：

$$SE_{\bar{y}_1-\bar{y}_0}=\hat{\sigma}\sqrt{\frac{2}{n}}$$

这就得到了对零假设$\bar{y}_1-\bar{y}_0=0$的检验（为了使检验统计量为正，对均值差异取绝对值）：

$$t=\frac{|\bar{y}_1-\bar{y}_0|}{\hat{\sigma}\sqrt{\frac{2}{n}}}$$

若使用效应值（方程3.10），其中$\delta=\frac{|\bar{y}_1-\bar{y}_0|}{\sigma}$，在非均衡设计中就得到：

$$t=\delta\sqrt{N\bar{T}(1-\bar{T})} \qquad [4.19]$$

若是均衡设计，则得到：

$$t=\delta\sqrt{\frac{n}{2}} \qquad [4.20]$$

利用BMI数据的检验实例

附录中的BMI数据表呈现了本章分析要用到的原始数据，结果变量（BMI）的概要统计参见表4.1。总的BMI均值$\bar{y}=33.4$，所有观测值的标准差$\sigma=5.997$。控制组均值$\bar{y}_0=35.340$，干预组的均值$\bar{y}_1=31.459$。控制组的标准差$\sigma_0=5.966$，干预组的标准差$\sigma_1=5.476$。[9]因此，总体标准差的估计值，合并标准差约为$\sqrt{(5.996^2+5.476^2)/2}=5.726$。[10]

均值差异用$\bar{y}_1-\bar{y}_0$进行估计，在BMI数据例子中，即$31.459-35.340=-3.881$。干预组的BMI均值比控制组的BMI均值低3.881。该例数据中，总体标准差的估计值（合并标准差）为$\sigma=5.726$。[11]因此得到效应值为：

$$\delta=\frac{|\bar{y}_1-\bar{y}_0|}{\sigma}=\frac{3.881}{5.726}=0.678$$

由于该数据属于均衡设计，每组包含 25 个观测值，因此检验统计量为[12]：

$$t=\delta\sqrt{\frac{n}{2}}=0.678\sqrt{\frac{25}{2}}=2.397$$

双尾检验的右侧临界值是自由度为 48 的 t 分布上 $1-0.025=0.975$ 处的分位数，等于 2.011。检验统计量的值(2.397)大于临界值 2.011，因此根据 $\alpha=0.05$ 水平上的双尾检验，我们拒绝零假设。

功效的非中心化参数

我们该如何根据这一检验结果来规划研究？通常我们无法确切知道结果变量的标尺单位。这就使预估均值差异和合并方差颇为困难。为了简化过程，我们可以对检验统计量的值进行标准化，得到一个无标尺的参数，如效应值(方程 3.10)，但我们现在用估计标准差替代效应值计算公式(方程 3.10)中的总体标准差。利用无标尺参数，我们可以在不掌握具体测量信息的情况下进行研究规划和设计。

运用 t 分布的功效分析不同于运用正态分布的功效分析。这是因为 t 分布的形状不同于非中心化曲线。图 4.1 展现了典型的 t 分布曲线，其非中心化参数分别是 $\lambda=0$，$\lambda=2$ 和 $\lambda=3$。因此，标准正态分布中“有用的算式关系”不能直接套用到这些样本中，但在下文我会说明这些算式关系仍然很有帮助。

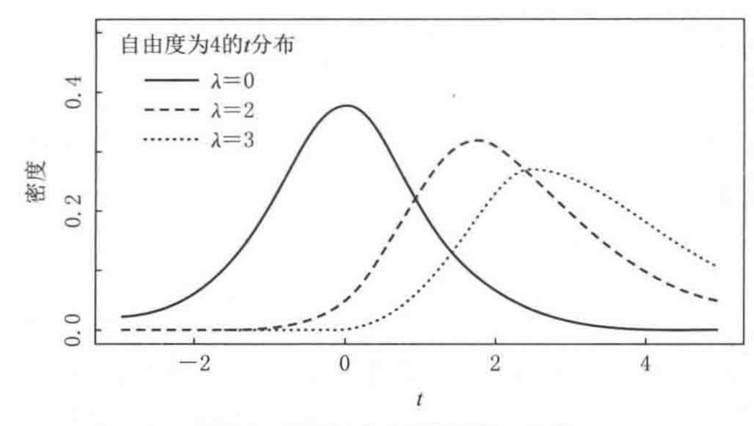


图 4.1 不同非中心化参数的 t 分布

假定实际的均值差异和总体方差已知,那么非中心化 t 分布的非中心化参数就是检验统计量的期望值。我们还是把这个值记作 λ。非均衡设计中,其表达式为:

$$\lambda = \underbrace{\delta}_{\text{效应值}} \underbrace{\sqrt{N\,\bar{T}(1-\bar{T})}}_{\text{样本量和分布比例}} \qquad [4.21]$$

均衡设计中,其表达式为:

$$\lambda = \underbrace{\delta}_{\text{效应值}} \underbrace{\sqrt{\frac{n}{2}}}_{\text{样本量}} \qquad [4.22]$$

我们将利用这些表达式来确定估计先验功效的策略,找出最低样本量和最低可检测效应值。

第 3 节 | 无协变量样本的功效分析

求先验功效

回顾第 3 章中检验的 β 参数（第二类错误）的计算，在单尾检验中其值等于备择（非中心化）分布中，虚无（中心化）分布的正临界值之前（或左侧）的面积，在双尾检验中则是备择分布曲线下虚无分布的两个临界值之间的面积。在单尾 t 检验中，β 参数的计算方程为：

$$\beta = \underbrace{H[t_{(df)1-\alpha},\ df,\ \lambda]}_{\text{右侧临界值之前（左侧）的面积}} \quad [4.23]$$

在双尾 t 检验中，β 参数的计算方程为：

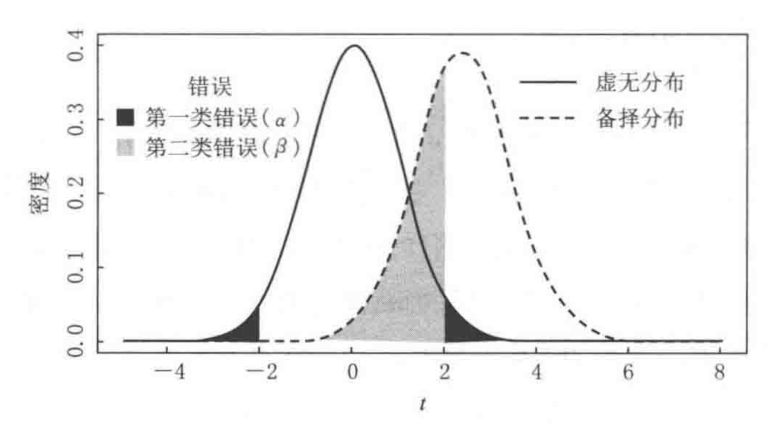
$$\beta = \underbrace{H[t_{(df)1-\alpha/2},\ df,\ \lambda]}_{\text{右侧临界值之前的面积}} - \underbrace{H[t_{(df)\alpha/2},\ df,\ \lambda]}_{\text{左侧临界值之前的面积}} \quad [4.24]$$

方程中的 $H[a,\ b,\ c]$ 表示自由度为 b、非中心化参数为 c 的非中心化 t 分布（备择分布）在 a 点的累积密度函数。$t_{(df)q}$ 是特定自由度的中心化 t 分布（虚无分布）上第 q 个分位数，即临界值。在任一情况下，功效都是上述面积的补集，即，

$$功效 = 1 - \beta \quad [4.25]$$

大多数统计软件包都提供这个 H 函数,虽然名称可能有所不同。[13]

功效是直接由第一类错误(α)和第二类错误(β)组成的一个函数。譬如,我们希望效应值 $\delta=0.678$,每组的样本量为 25,那么检验统计量预计约为 $t=2.397$,其自由度为 $2n-2=48$。这意味着非中心化参数(λ)也约为 2.397。图 4.2 对这一分析进行了图形化展示,与正态分布例子中的流程非常相似。我们得到这一检验的功效为 0.651。



注:$\alpha=0.05$,自由度为 48 的双尾检验,因此临界值为 2.011。检验统计量等于 2.397,因此 $\beta=0.349$(灰色阴影面积),得到功效为 0.651。

图 4.2　BMI 数据功效分析结果

求样本量

研究者经常希望知道,为了检测到某效应,所需要的最低样本量是多少。遗憾的是,这无法通过对方程 4.23 或方程 4.24 进行变换得到,因为累积密度函数本身在自由度参数中包含了样本量(参见方程 4.24)。因此,必须首先利用标准正

态分布求近似值，如方程 3.9，因为它不需要自由度。然后利用前面的自由度结果对分析进行修正。

回顾第 3 章的内容，对于总体方差已知的正态分布变量，预计的均值差异检验与 z 分布上分位数的关系如方程 3.9 所示，$\lambda \approx z_{critical} - z_{\beta}$，从而在非均衡设计中得到方程 3.13：

$$N \approx \frac{(z_{critical} - z_{\beta})^2}{[P_1(1-P_1)]\delta^2}$$

在均衡设计中得到方程 3.14：

$$n \approx \frac{2(z_{critical} - z_{\beta})^2}{\delta^2}$$

虽然很简单，但必须牢记我们正在处理的样本服从 t 分布。因此，利用标准正态分布求近似值只是第一步，得到的近似值往往小于实际所需的最低样本量。这是因为方程 3.9 是基于正态曲线的，而正态分布的分位数值总是小于 t 分布的分位数值。

解决办法是以标准正态分布的近似值（方程 3.13 和方程 3.14）为第一步，得到一个初始样本量，然后利用初始样本量求基于自由度的 t 分布分位数近似值：

$$\lambda \approx t_{(N_z-2)critical} - t_{(N_z-2)\beta} \qquad [4.26]$$

我使用 N_z-2 或 $2n_z-2$ 是为了提醒读者，这是源自标准正态分布近似值的自由度。

然后我们在新公式中使用这一近似值，以求得最低样本量。在非均衡设计中，这一计算公式如下：

$$N \approx \frac{[t_{(N_z-2)critical} - t_{(N_z-2)\beta}]^2}{[\bar{T}(1-\bar{T})]\delta^2} \qquad [4.27]$$

在均衡设计中，计算公式如下：

$$n \approx \frac{2[t_{(2n_z-2)critical} - t_{(2n_z-2)\beta}]^2}{\delta^2} \qquad [4.28]$$

这些步骤得到的仍是不精确的结果，因此一般而言，最好先用上述分析得到的样本量计算先验功效以进行核查。随着样本量的变大，近似值的误差水平自然会减小，因为当自由度大于 100 之后，t 分布就非常接近标准正态分布了。

最好是利用计算机来进行大多数的精确运算。然而，作为学习工具，附录表 A.4 也提供了 t 分布的分位数值，对应的自由度取值分别是 2—25，以及 25—100 之间 5 的倍数。我们会在实例中用到这些取值，因此就不一定需要计算机的帮助。例如，为了求 $\alpha=0.05$ 且自由度为 10 的双尾检验临界值，我们找到附录表 A.4 中第一列取值为 10 的那一行，然后找到 $1-\alpha/2=1-0.05/2=1-0.025=0.975$ 的分位数，得到的临界值为 2.228。为了求功效为 0.8、自由度为 10 的分位数，我们找到同一行的 $1-0.8=0.2$ 的分位数，即为 -0.879。

非均衡设计实例

假定我们计划进行某个双尾检验，$\alpha=0.01$，效应值 $\delta=2$，其中 35%的样本属于组 1，因而 $\bar{T}=0.35$。使用标准正态分布的步骤（方程 3.13）和分位数表求出功效为 0.9（$\beta=0.1$）时的样本量：

$$N_z \approx \frac{(Z_{0.995}-Z_{0.1})^2}{[\bar{T}(1-\bar{T})]\delta^2} \approx \frac{[2.576-(-1.282)]^2}{[0.35(1-0.35)]2^2} \approx 16.356$$

即约为 17 个观测值。

接下来，我们使用分位数表求自由度为 17－2＝15 的 t 分布在 0.995 和 0.1 的分位数，分别为 2.947 和－1.341。把这些值代入方程 4.27，就得到样本量为：

$$N \approx \frac{[t_{(N_z-2)critical} - t_{(N_z-2)\beta}]^2}{[\bar{T}(1-\bar{T})]\delta^2} \approx \frac{[2.947-(-1.341)]^2}{[0.35(1-0.35)]2^2} \approx 20.205$$

即约为 21 个观测值，其中一组约为 8 个，另一组 13 个。这一设计的实际功效是 0.930，可见我们的近似值提供的功效比预计的要多。

均衡设计实例

现在考虑跟上面例子一样的分析，$\alpha=0.01$，效应值 $\delta=2$，只不过现在假定是均衡设计。为了求功效为 0.9 时每个组的样本量，我们第一步还是使用正态分布的计算步骤：

$$n_z \approx \frac{2(z_{critical} - z_\beta)^2}{\delta^2} \approx \frac{2[2.576-(-1.282)]^2}{2^2} \approx 7.442$$

即每组约 8 个个案。

接下来，我们使用分位数表求自由度为 14（因为 2×8－2＝14）的 t 分布在 0.995 和 0.1 的分位数，分别为 2.997 和－1.345。把这些值代入方程 4.28，得到样本量为：

$$n \approx \frac{2[t_{(2n_z-2)critical} - t_{(2n_z-2)\beta}]^2}{\delta^2} \approx \frac{2[2.977-(-1.345)]^2}{2^2} \approx 9.340$$

即约每组 10 个观测值，共 20 个观测值。这一设计的实际功效为 0.929。

求最低可检测效应

假定已经确定了样本量,但研究者希望知道特定 α 水平和功效$(1-\beta)$水平下该研究的最低可检测效应值(MDES)(Bloom, 1995)。这个值是某个设计在既定功效水平和检验情况下能检测到的最小效应值,标记为 δ_m。犹如处理标准正态分布一样,我们可以根据特定检验类型和 α 水平下的 t 分布分位数,以及所要求的第二类错误(β)来计算 MDES。非均衡设计中的计算式如下:

$$\delta_m \approx \frac{t_{(N-2)critical} - t_{(N-2)\beta}}{\sqrt{N\bar{T}(1-\bar{T})}} \qquad [4.29]$$

而均衡设计中的计算式如下:

$$\delta_m \approx [t_{(2n-2)critical} - t_{(2n-2)\beta}]\sqrt{2/n} \qquad [4.30]$$

需要注意的是,这里不需要使用标准正态分布近似值,因为我们已经设定了样本量,所以自由度已知。为了与上文的样本量标记保持一致,我这里还是用 $N-2$ 和 $2n-2$ 来分别标记非均衡设计和均衡设计中的自由度。

非均衡设计实例

假定正计划进行某个检验,其中 20%的个案属于干预组,故而$\bar{T}=0.2$,在 $\alpha=0.05$ 且功效为 0.8(即 $\beta=0.2$)水平下进行单尾检验,我们想知道共有 82 个个案的情况下最低可检测效应是多少。已知自由度为 $N-2=80$,利用分位数表可得 $t_{(2n-2)critical}=t_{(80)1-0.05}=t_{(80)0.95}=1.664$,且 $t_{(2n-2)\beta}=-0.846$。因而,MDES 的计算过程如下:

$$\delta_m \approx \frac{t_{(N-2)critical} - t_{(N-2)\beta}}{\sqrt{N\,\bar{T}(1-\bar{T})}} \approx \frac{1.664-(-0.846)}{\sqrt{82\times 0.2(1-0.2)}} \approx 0.693$$

这说明该设计可检测到的效应值为0.693个标准差单位。

均衡设计实例

假定正计划进行$\alpha=0.05$且功效为0.8(即$\beta=0.2$)水平下的单尾检验，我们想知道共有82个个案的情况下最低可检测效应是多少。这次样本是均衡的，每组包含41个个案。已知自由度为$2n-2=80$，还是利用分位数表，得到$t_{(2n-2)critical}=t_{(80)1-0.05}=t_{(80)0.95}=1.664$，且$t_{(2n-2)\beta}=-0.846$。因而，MDES的计算过程如下：

$$\begin{aligned}\delta_m &\approx [t_{(2n-2)critical} - t_{(2n-2)\beta}]\sqrt{2/n} \\ &\approx [1.664-(-0.846)]\sqrt{2/41} \approx 0.554\end{aligned}$$

这说明该设计可检测到的效应值为0.554个标准差单位。

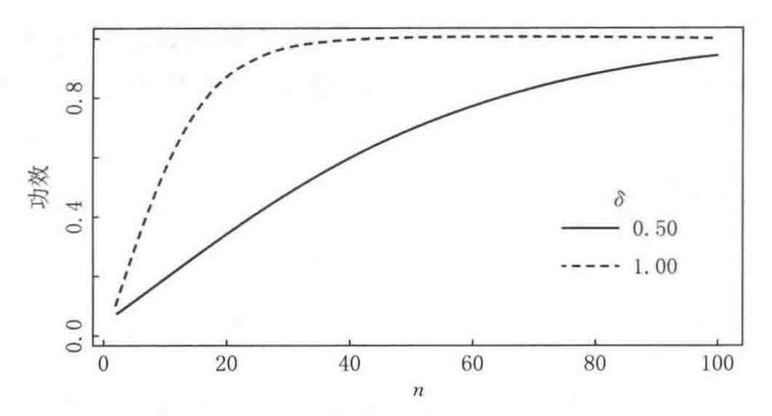
对功效的影响

通过上述实例，我们可以看到很多参数都会影响功效。若检验类型(通常为双尾检验)、显著性水平(通常为$\alpha=0.05$)和均衡性(如果设计一个实验，我们通常假定均衡设计以最大化功效)都保持不变，我们可以通过画图来更好地理解效应值、样本量和功效之间的关系。

效应值对功效有很大影响。如图4.3所示，假定$\alpha=0.05$水平下的双尾检验，功效总是随着样本量(n)的增加而增加。然而，对低效应值而言，这一曲线更为平缓。譬如，即使每组

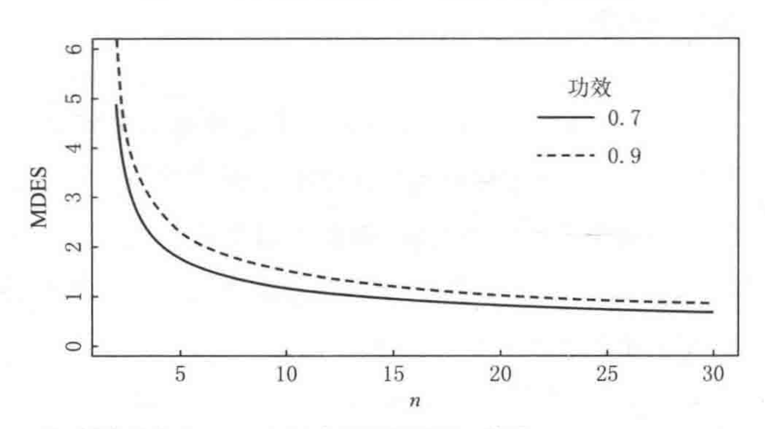
都有 40 个观测值，效应值为 0.5 时功效仍然颇低。相反，若效应值较大，如等于 1，能在每组观测值都较少的情况下获得足够功效。

反之亦然，功效和样本量也会影响最低可检测效应值。如图 4.4 所示，假定 $\alpha=0.05$ 水平下的双尾检验，可能检测到的效应值总是随着样本量(n)增加而降低。图中的两条线告诉我们，功效越低，可检测效应值也越低。



注：样本量为 $2n$，$\alpha=0.05$ 水平下的双尾 t 检验。

图 4.3　不同效应值(δ)和 n 情况下的功效曲线



注：样本量为 $2n$，$\alpha=0.05$ 水平下的双尾 t 检验。

图 4.4　不同功效和 n 情况下的最低可检测效应值

第 4 节 | 小结

本章考察了如何对简单随机样本中的两组均值差异检验进行功效分析。首先，我们运用回归分析讨论了进行这一检验的基本方法。然后定义了功效分析的参数，探讨了均衡设计与非均衡设计中的不同计算方法。由于我们发现非均衡设计会导致更低的功效，因此在接下来的论述中我们主要就均衡设计的分析进行讨论。

下一章将引入协变量。我们还是利用回归分析来理解协变量对组均值差异检验的效应。

第5章

在均衡设计的简单组均值差异检验中引入协变量

本章将对组均值差异分析进行拓展，引入协变量的使用。某个协变量对功效的影响可能为正，也可能为负，这取决于干预指示变量与协变量之间的关系。本章将概述两种情况下的功效参数：一是干预指示变量与协变量相关；二是由于干预条件的随机化，干预指示变量与协变量无关。为行文简便起见，我们仅考虑均衡设计的情况。

由于本章将深入细致地讨论检验及其参数，对两组设计中的协变量使用进行概念化处理或许会有所裨益。切记，统计功效是基于统计检验的，而后者又基于效应的抽样方差。因此，任何使抽样方差变小的因素都会导致检验更有功效。

当有可能找到与结果变量相关而与其他预测变量不相关的变量时，把这些变量纳入回归模型中就能减小残差方差。而缩减残差方差能改善功效。回顾一下回归系数的抽样方差计算公式（方程 4.12）：

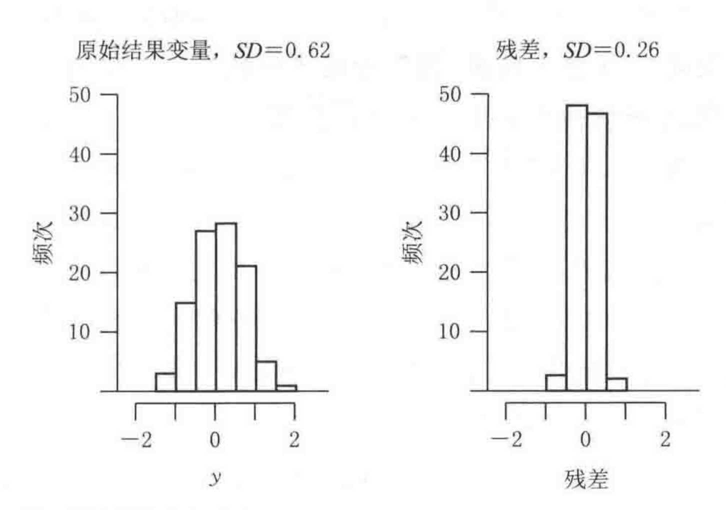
$$\hat{\text{var}}\{\hat{\gamma}_1\} = \frac{\hat{\sigma}_e^2}{\sum_j \sum_i (T_{ij} - \bar{T})^2}$$

这一表达式说明，回归斜率的抽样方差（用于获取标准误）是基于残差（$\hat{\sigma}_e^2$）的方差计算得到的。残差方差越小，系数的抽

样方差就越小。模型纳入的协变量越多，残差方差就会越小。只要预测变量之间不相关，那么纳入协变量就一定会使抽样方差变小。

可以对回归如何带来更小的抽样方差进行图形化处理，参见图 5.1 中的两个直方图。左侧是某个任意变量的直方图。接下来，右侧图是左侧直方图中变量与另一任意预测变量进行回归后的残差直方图。如你所见，右侧的残差方差（散布情况）小于左侧的原始结果变量。

在下一部分我们将会看到，由于预测变量之间存在相关，有时候加入协变量并不能改善功效。



注：残差的标准差更小。

图 5.1 某随机结果变量及其回归模型残差的直方图和标准差

第 1 节 | 实例分析

我们首先回到 BMI 数据。回顾一下第 4 章的内容,比较吃早餐和不吃早餐两组人时,我们发现两组的 BMI 分数存在显著差异。表 5.1 的"仅有干预变量"回归部分重复了这一结果。在第二个回归模型中,我们增加了一个研究初期做的前测变量。"干预+前测"模型报告了这些结果。在第二个模型中,干预效应小了很多,且不再显著。

表 5.2 的前测概要统计揭示了干预效应变小且不显著的原因。我们看到指派到控制组的 BMI 平均值要高于那些指派到干预组的 BMI 均值。这意味着干预指示变量与协变量之间存在相关,这就导致了多重共线性问题,从而减小了功效。当随机化有效时,干预指示变量和协变量之间就不会存在相关,这时增加协变量就能明显改善功效。

表 5.1 预测 BMI 的模型

	仅有干预变量	干预+前测
截距	35.340(1.145)***	−0.617(0.992)
吃早餐对比不吃早餐	−3.881(1.620)*	−0.084(0.315)
前测		1.017(0.027)***
R^2	0.107	0.970
N	50	50
均方误差根(Root MSE)	5.727	1.052

注:BMI 为体质指数;MSE 为均方误差。括号中为标准误。* $p<0.05$, ** $p<0.01$, *** $p<0.001$。

表 5.2　前测体质指数的概要统计

	N	均值	标准差
不吃早餐(控制组)	25.000	35.360	5.586
吃早餐(干预组)	25.000	31.626	5.489
总样本	50.000	33.493	5.796

第 2 节 | 均衡样本中运用协变量的检验

很多情况下，研究者都希望在组均值差异分析中包含协变量。这就是所谓的协变量分析（ANCOVA）。协变量分析主要是为了增加随机试验的精确性，但也可用于准实验中以消除虚假效应，或直接应用于对某些因素的效应感兴趣但又想增加控制变量的观测性研究（Wildt & Ahtola，1978）。

然而，就功效分析而言，运用协变量可能会带来风险，因为干预指示变量与协变量之间的任何相关都会改变均值差异和抽样方差的估计值。一般而言，这会导致检验统计量变小。但如若能确保干预指示变量与协变量之间相关接近于0，且协变量与结果变量之间的相关程度较高，那么加入协变量就会对功效有较大的改善作用。

定义检验

我们从数据产生过程开始，在方程 4.1 中增加一个新变项：

$$y_{ij} = \mu + \tau_j' + \phi_1 (x_{ij} - \bar{x}) + e_{ij} \qquad [5.1]$$

其中 τ'_j 是修正干预效应，ϕ_1 是预测结果变量 y 的协变量 x 的组内斜率。[14] 利用组内效应回归模型(即计量经济学家所说的固定效应模型)可对 ϕ_1 进行估计：

$$y_{ij}-\bar{y}_j+\bar{y}=\hat{\phi}_0+\hat{\phi}_1(x_{ij}-\bar{x}_j+\bar{x})+r_{ij} \qquad [5.2]$$

由方程 5.1 可知，左侧结果变量可通过协变量的变换而得到类似于方程 4.1 的形式：

$$y_{ij}-\phi_1(x_{ij}-\bar{x})=\mu+\tau'_j+e_{ij} \qquad [5.3]$$

这意味着修正干预效应 τ'_j 通过 ϕ_1 及协变量在干预组和控制组之间的均值差异，从而与未修正干预效应相关(Porter & Raudenbush，1987)：

$$\tau'_j=\tau_j-\phi_1(\bar{x}_1-\bar{x}_0) \qquad [5.4]$$

当 T 和协变量 x 之间不相关时，x 在每组上的均值相等，换言之，$\tau'_j=\tau_j$。

回归中的协变量分析

在这一部分，我们从多元回归的框架来思考检验统计量，这允许我们把非中心化参数分解为效应值和相关系数。

我们在前面用线性回归模型来估计均值差异(方程 4.7)。在这一部分，我们在回归模型中增加了均值对中(mean-centered)的协变量：

$$y_{ij}=\gamma_0+\gamma_1 T_{ij}+\phi_1(x_{ij}-\bar{x})+e_{ij} \qquad [5.5]$$

对该模型的最小二乘估计同时估计了修正均值差异和 x 对 y 的组内效应。这对均值差异的估计有所影响。修正均值差

异不再是$\hat{\gamma}_1=\bar{y}_1-\bar{y}_0$,而是变成了:

$$\hat{\gamma}_1=(\bar{y}_1-\bar{y}_0)-\phi_1(\bar{x}_1-\bar{x}_0) \qquad [5.6]$$

接下来,修正均值差异的抽样方差也不再是$\sigma^2\frac{2}{n}$。这是因为加入协变量之后,σ^2 的估计值不再等于均方误差的估计值σ_e^2。回归均方误差的估计值现在是(Kirk, 1995):

$$\hat{\sigma}_e^2=\hat{\sigma}^2(1-\hat{\rho}_{yx,w}^2)\frac{pn-p}{pn-p-1} \qquad [5.7]$$

其中 $\rho_{yx,w}$为结果变量 y 和协变量 x 之间的组内相关,估计值为[15]:

$$\hat{\rho}_{yx,w}=\frac{\sum_j\sum_i(y_{ij}-\bar{y}_j)(x_{ij}-\bar{x}_j)}{\sqrt{\sum_j\sum_i(y_{ij}-\bar{y}_j)^2\sum_j\sum_i(x_{ij}-\bar{x}_j)^2}} \qquad [5.8]$$

而$\frac{pn-p}{pn-p-1}$是对合并标准差的一个必要修正系数,因为协变量消耗了一个自由度。$\rho_{yx,w}$的最好阐释是,在假定没有干预或组效应情况下 y 和 x 之间总体相关的估计值。而 y 和 x 之间的样本相关不是一个好的估计值,因为已引入了对 y 的干预效应。

大家都知道多元回归中任何变量 q 对应斜率的抽样方差是(参见 Fox, 2015):

$$\text{var}\{\gamma_q\}=\frac{\sigma_e^2}{\sum(x_q-\bar{x}_q)^2}\frac{1}{1-\rho_{qx}^2} \qquad [5.9]$$

其中 σ_e^2 是回归的均方误差,$\sum(x_q-\bar{x}_q)^2$ 是预测变量的平

方和，$\frac{1}{1-\rho_{qx}^{2}}$是由于预测变量 q 和其他协变量 x 之间的多元相关导致的方差膨胀因子（VIF）。方差膨胀因子是增加回归效应抽样方差的一个因素，其大小与预测变量和其他协变量之间的相关强度存在对应关系。预测变量之间的相关越强，该因素导致抽样方差的膨胀效应就越大。对于均衡样本中的预测变量 T 而言，$\sum_j \sum_i (T_{ij} - \bar{T})^2$ 等于 $n/2$（方程 4.16）。在本例中，我们感兴趣的系数是 γ_1，即两组规模相等、存在协变量 x 的情况下 T 对 y 的效应，该表达式就变成：

$$\mathrm{var}\{\gamma_1\} = \sigma_e^2 \frac{2}{n} \frac{1}{1-\rho_{Tx}^2} \qquad [5.10]$$

因此，若以方程 5.7 替换 σ_e^2，该抽样方差（假定两组样本量相等）就变成：

$$\widehat{\mathrm{var}}\{\hat{\gamma}_1\} = \hat{\sigma}^2(1-\hat{\rho}_{yx.w}^2)\frac{2n-2}{2n-3}\frac{2}{n}\frac{1}{1-\hat{\rho}_{Tx}^2} \qquad [5.11]$$

其中 $\hat{\sigma}^2$ 为根据方程 4.6 得到的总体方差估计值，并用 $(1-\hat{\rho}_{yx.w}^2)$ 进行了向下调整。最后，$\frac{1}{1-\hat{\rho}_{Tx}^2}$ 是方差膨胀因子，来自干预指示变量与协变量之间相关系数的平方，或者说 T 的方差中被 x 解释掉的部分。

BMI 实例数据的分析

检验系数 γ_1 的流程之第一步是估计 y 对 x 回归的组内斜率、y 和 x 的组内相关、干预指示变量和协变量之间的相

关。在 BMI 实例数据中,这三个估计值分别是 $\hat{\phi}_1=1.017$,$\hat{\rho}_{yx.w}=0.983\,3$,$\hat{\rho}_{Tx}=-0.325$。

根据方程 5.6,利用表 4.1 和表 5.2 中的值以及 $\hat{\phi}_1$,可以计算得到数据中的修正组间差异:$\hat{\gamma}_1=-3.881-1.017\times(31.626-35.360)=-0.084$。

为了计算这一差异的抽样方差估计值,运用方程 5.11 得到:

$$\hat{\text{var}}\{\hat{\gamma}_1\}=5.727^2(1-0.983\,3^2)\frac{50-2}{50-3}\frac{2}{25}\frac{1}{1-(-0.325)^2}=0.099$$

其平方根为 0.315,就是表 5.1 中的标准误。

最后,进行 t 检验运算,$t=\frac{|-0.084|}{0.315}=0.267$,已知自由度为 47,则双尾检验的 p 值为 0.605。这一 p 值离统计显著的标准颇远。下一节的讨论将对上文的论述和计算进行阐释。

定义功效参数

回归分析框架能整理得到非中心化参数,因为可以参照未纳入协变量的情况,以斜率来构建效应值。利用方程 5.6 和方程 5.11,可得到非中心参数为:

$$\lambda=\frac{(\mu_1-\mu_0)-\phi_1(\bar{x}_1-\bar{x}_0)}{\sqrt{\sigma^2(1-\rho^2_{yx.w})\frac{2n-2}{2n-3}\frac{2}{n}\frac{1}{1-\rho^2_{Tx}}}}$$

上式可变换为乘积的形式,如下式所示,依次分别包含修正效应值、协变量效应倒数的平方根、样本量和多重共线性效

应四个乘积项：

$$\lambda=\underbrace{\frac{(\mu_1-\mu_0)-\phi_1(\bar{x}_1-\bar{x}_0)}{\sigma}}_{\text{修正效应值}}\underbrace{\sqrt{\frac{1}{(1-\rho^2_{yx,w})}}}_{\text{协变量效应}}\underbrace{\sqrt{\frac{n}{2}\frac{2n-3}{2n-2}}}_{\text{样本量}}\underbrace{\sqrt{1-\rho^2_{Tx}}}_{\text{多重共线性}} \quad [5.12]$$

没有协变量的功效分析(方程 4.22)只有两个参数，即效应值和样本量。反之，如方程 5.12 所示，存在相关协变量的功效分析则要求更多的先验信息，包括协变量对结果变量变异($\rho^2_{yx,w}$)之解释的有效性，以及分组指示变量与协变量之间可能存在相关(ρ_{Tx})的一些信息。修正效应值的期望是：

$$\delta_a=\frac{(\mu_1-\mu_0)-\phi_1(\bar{x}_1-\bar{x}_0)}{\sigma} \quad [5.13]$$

样本量也可以被简化。系数$\frac{2n-3}{2n-2}$总是小于 1，但当样本量达到 $n=30$ 时，该系数趋近于 1。然而，当样本量 $n\geqslant 6$ 时，这一表达式的平方根永远大于 0.95。因此，除非样本量非常小，否则这一系数对非中心化参数的影响非常小。掌握了这一点和效应值，那么方程 5.12 就约等于下述简化表达式：

$$\lambda\approx\delta_a\sqrt{\frac{n}{2}}\sqrt{\frac{1}{1-\rho^2_{yx,w}}}\sqrt{1-\rho^2_{Tx}} \quad [5.14]$$

第 3 节 | 协变量与干预指示变量相关情况下的功效分析

当协变量与干预指示变量相关时,功效分析就变得颇为麻烦。干预指示变量和协变量之间的相关性质甚至比效应值更难预估。

预估干预变量和协变量之间的相关

猜测干预变量与协变量间相关的一种可能方法是给协变量假定一个效应值。已知两个变量之间的相关是斜率与标准差比率的乘积,$\rho_{ab}=\gamma_{ab}\frac{\sigma_b}{\sigma_a}$。由于变量 a 和 b 的相关等于 b 和 a 的相关,而且在组样本量均衡的情况下,T 预测 x 的斜率是$\bar{x}_1-\bar{x}_0$,我们就可以把 ρ_{Tx} 表示为:

$$\rho_{Tx}=(\bar{x}_1-\bar{x}_0)\frac{\sigma_T}{\sigma_x}$$

若 T 为均衡分组的二分指示变量,上式就可以进一步表示为:

$$\rho_{Tx}=\frac{(\bar{x}_1-\bar{x}_0)}{\sigma_x}\sqrt{\frac{pn}{4(pn-1)}}=\delta_{\bar{x}_1-\bar{x}_0}\sqrt{\frac{pn}{4(pn-1)}}$$

其中 $\delta_{\bar{x}_1-\bar{x}_0}$ 为效应值，测量了干预组和控制组在协变量上的标准化差异，这意味着在均衡大样本中，得到 ρ_{Tx}^2 的近似值如下：

$$\rho_{Tx}^2 \approx \frac{\delta_{\bar{x}_1-\bar{x}_0}^2}{4} \qquad [5.15]$$

譬如，表 5.2 中，干预组和控制组在协变量上的差异是 31.626－35.360＝－3.734，然后再除以 x 的总标准差，得到 $\delta_{\bar{x}_1-\bar{x}_0}=-3.734/5.796=-0.644$。若对这一效应值进行平方，其值为 0.415，再除以 4，得到 0.104。这就是 ρ_{Tx}^2 的近似值。[16] 值得注意的是，本例表明仅有 10%的协变量方差能被干预指示变量加以解释，这一效应并不大。但这足以使分析有偏，并消除效应的显著性。

求先验功效

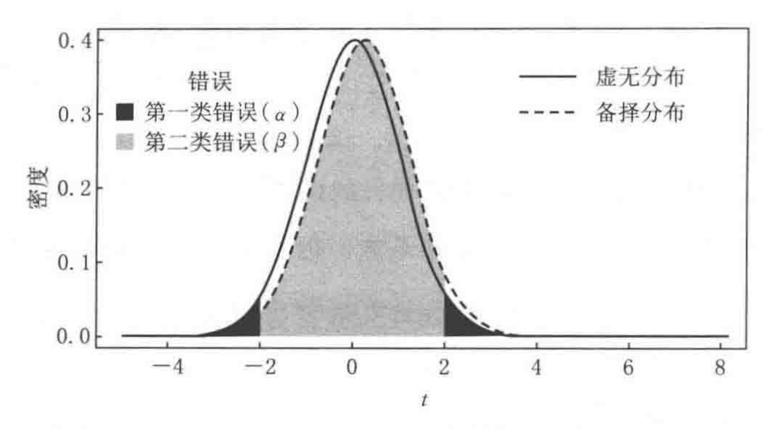
若忽略效应的方向，上文实例数据分析中的修正效应值（方程 5.13）约为 $|\hat{\delta}_a|=0.084/5.727=0.014\,7$。假定存在某个协变量，其与结果变量的组内相关为 $\rho_{yx.w}=0.983\,3$，其与干预指示变量的总相关为 $\rho_{Tx}=-0.325$。鉴于效应值较低，且干预变量与协变量存在相关，因而得到一个较小的 t 统计量 $|t|=0.267$。这也等于 λ 的值，

$$\lambda=\frac{(\mu_1-\mu_0)-\phi_1(\bar{x}_1-\bar{x}_0)}{\sigma}\sqrt{\frac{1}{(1-\rho_{yx.w}^2)}}\sqrt{\frac{n}{2}\frac{2n-3}{2n-2}}\sqrt{1-\rho_{Tx}^2}$$

$$\lambda=0.014\,7\sqrt{\frac{1}{(1-0.983\,3^2)}}\sqrt{\frac{25}{2}\frac{50-3}{50-2}}\sqrt{1-(-0.325)^2}=0.267$$

根据方程 4.24,得到功效约为 0.058(参见图 5.2)。

这再一次说明,当存在相关的协变量时,功效分析必须处理(修正)效应值和样本量之外的几个参数,包括结果变量与协变量的相关系数($\rho_{yx.w}$)以及干预指示变量与协变量的相关系数(ρ_{Tx})。这两个相关系数对功效的作用方向正好相反。



注:$\delta_a=0.014\,7$, $\lambda=0.267$, $\beta=0.942$,功效为 0.058。

图 5.2 纳入协变量的情况下,BMI 实验结果的功效分析

如图 4.3 所示,功效随样本量的增加而增加。然而,样本量既定的情况下,功效随着干预指示变量与协变量相关系数的增加而减小。同干预指示变量与协变量相关系数的影响不同,样本量既定的情况下,功效随着结果变量与协变量相关系数的增加而增加。

图 5.3 展现了干预指示变量与协变量相关系数对功效的影响,假定 $\delta_a=0.50$, $\rho_{yx.w}=0.75$, $\alpha=0.05$ 的双尾检验。与图 4.3 所示相同,功效随样本量的增加而增加。然而,样本量既定的情况下,功效随着干预指示变量与协变量相关系数的

增加而减小。例如，当 n 为 40 时，干预指示变量与协变量的相关约为 0.25，修正效应值约为 0.5，结果变量与协变量相关系数约为 0.75，该检验的功效约为 0.9。当干预指示变量与协变量相关系数增加到 0.5 时，功效会受到一些影响。若干预指示变量与协变量相关系数为 0.75，对功效的影响颇大，功效降低到约为 0.6。

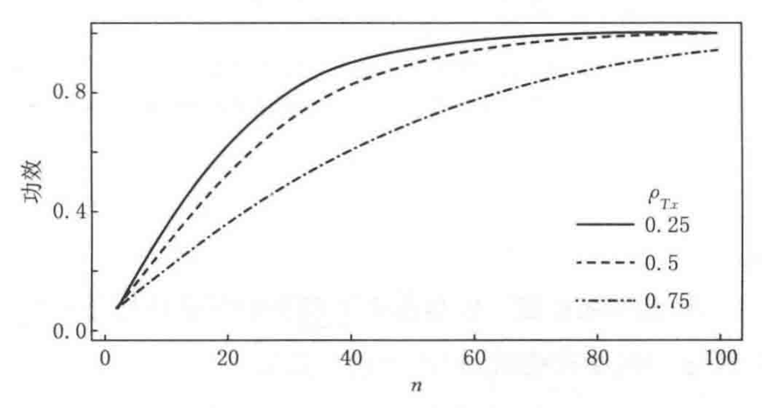


图 5.3　不同 n 和 ρ_{Tx} 的功效曲线
($\alpha=0.05$，$\rho_{yx,w}=0.75$，$\delta_a=0.50$ 的双尾 t 检验)

图 5.4 展现了结果变量与协变量相关系数对功效的影响，假定 $\delta_a=0.50$，$\rho_{Tx}=0.50$，$\alpha=0.05$ 的双尾检验。不同于干预变量与协变量相关系数对功效的影响，样本量既定的情况下，功效随着结果变量与协变量相关系数的增加而增加。例如，当 n 为 40 时，结果变量与协变量的相关系数约为 0.25，修正效应值约为 0.5，干预指示变量与协变量相关系数约为 0.5，该检验的功效约为 0.5。当结果变量与协变量的相关系数增加到 0.5 时，功效会受到一些影响。然而，若结果变量与协变量的相关系数为 0.75 时，就会对功效产生较大影响，功效会增加到 0.8 以上。

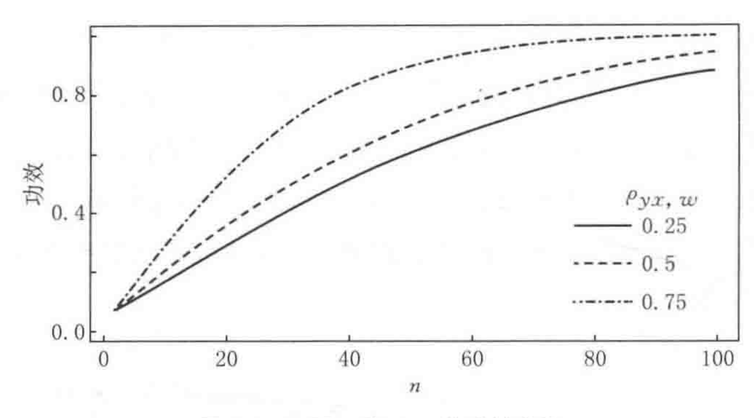


图 5.4 不同 n 和 $\rho_{yx,w}$ 的功效曲线
($\alpha=0.05$, $\rho_{Tx}=0.5$, $\delta_a=0.5$ 的双尾 t 检验)

图 5.5 展现了干预指示变量与协变量相关,以及结果变量与协变量相关对功效的联合效应,假定 $n=40$, $\delta_a=0.50$, $\alpha=0.05$ 的双尾检验。根据图中干预指示变量与协变量相关系数(ρ_{Tx})的变化曲线,相比于相关系数为 0 时的功效,随着相关系数的增加,功效迅速降低。而根据图中结果变量与协变量的相关系数($\rho_{yx,w}$)变化曲线,当系数大于 0.4 后,功效

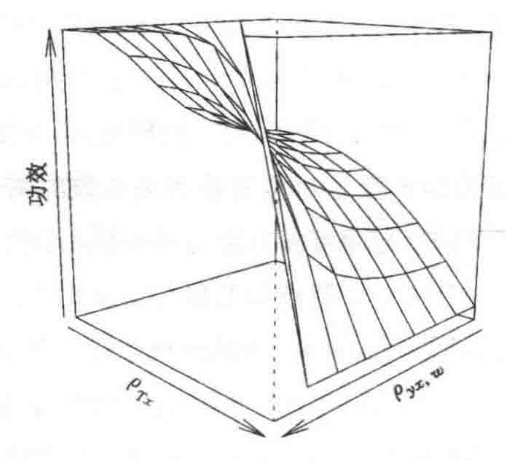


图 5.5 不同 ρ_{Tx} 和 $\rho_{yx,w}$ 的功效曲线
($\alpha=0.05$, $n=40$, $\delta_a=0.5$ 的双尾 t 检验)

迅速增加。由此可见，结论是尽量避免干预指示变量与协变量之间的相关。然而，若协变量与结果变量达到中度相关（大于 0.4），则使用协变量有利于提升功效。

求样本量

假定先验知识提供了协变量、结果变量和干预指示变量间相关的情况，我们可以利用方程 3.9 对方程 5.14 进行变换，得到近似的样本量方程：

$$n_Z \approx \frac{2(z_{critical} - z_\beta)^2}{\delta_a^2} \frac{1-\rho_{yx,w}^2}{1-\rho_{Tx}^2} \qquad [5.16]$$

上式提供了起始样本量的自由度，可应用到下式中 * ：

$$n \approx \frac{2[t_{(2n_Z-3)critical} - t_{(2n_Z-3)\beta}]^2}{\delta_a^2} \frac{1-\rho_{yx,w}^2}{1-\rho_{Tx}^2} \qquad [5.17]$$

这一表达式提供了很多信息，不仅表明样本量会随着效应值（δ_a）的增加而增加，也说明如果干预指示变量和协变量（ρ_{Tx}）存在相关，那么效应值对样本量的缩减效应会有所抵消。最后，当 y 和 x 的相关（$\rho_{yx,w}$）增加，样本量也会缩减。

实例

譬如，假定我们希望求某个均衡设计中每组的样本量，以得到修正效应值 $\delta_a = 0.7$，其中结果变量和协变量的相关

---

* 由于使用了一个协变量，所以又少了一个自由度，因此方程 5.17 和方程 5.21 中 t 值对应的自由度是 $2n_Z-3$，方程 5.18 和方程 5.22 中 t 值对应的自由度是 $2n-3$。原文有误，译者进行了勘误。——译者注

约为 $\rho_{yx,w}=0.8$，但实验组和协变量之间的相关较小，$\rho_{Tx}=0.1$。首先利用标准正态分布的分位数确定一个近似样本量。若假定 $\alpha=0.05$ 的双尾检验，希望获得 0.8 的功效，则标准正态分布上有用的分位数为 $z_{0.975}=1.96$ 和 $z_{0.2}=-0.842$，因此样本量近似值为：

$$n_z\approx\frac{2[1.96-(-0.842)]^2}{0.7^2}\frac{1-0.8^2}{1-0.1^2}\approx11.653$$

即每组约为 12 个个案。然后选择合适的 t 分布分位数，自由度为 $2\times12-3=21$，得到 $t_{(21)0.975}=2.08$，$t_{(21)0.2}=-0.859$，由此得到：

$$n\approx\frac{2[2.08-(-0.859)]^2}{0.7^2}\frac{1-0.8^2}{1-0.1^2}\approx12.820$$

即每组约为 13 个个案。

求最低可检测效应

再次对方程 5.14 进行变换，得到最低可检测的修正效应值：

$$\delta_{a,m}\approx[t_{(2n-3)critical}-t_{(2n-3)\beta}]\sqrt{\frac{2}{n}}\sqrt{\frac{1-\rho_{yx,w}^2}{1-\rho_{Tx}^2}} \qquad [5.18]$$

这一表达式提供了很多信息，因为它说明了两个相关系数的相对效应，即结果变量与协变量之间相关和干预变量与协变量之间相关的效应之相对大小。结果变量与协变量之间相关的效应翻倍了，而干预变量与协变量之间相关的效应则要再乘以组样本量。因此，假定两个相关系数相等，那么随着

样本量的增加，干预变量与协变量之间相关的效应要远大于结果变量与协变量之间相关的效应。

实例

根据方程 5.18，我们还是用 t 分布的分位数来求最低可检测效应。假定进行 $\alpha=0.05$ 的双尾检验，功效为 0.8（因此 $\beta=0.2$），我们设计了一个均衡研究，每组样本量为 $n=24$，因此自由度为 $2n-3=45$。我们还预计结果变量和协变量之间的相关约为 $\rho_{yx,w}=0.8$，而干预组和协变量之间的相关较小，即 $\rho_{Tx}=0.1$。由于 $t_{(45)0.975}=2.014$，$t_{(45)0.2}=-0.85$，因而该设计的最低可检测修正效应值约为 0.499。

$$\delta_{a,m} \approx [2.014-(-0.85)]\sqrt{\frac{2}{24}}\sqrt{\frac{1-0.8^2}{1-0.1^2}} \approx 0.499$$

第4节 协变量与干预指示变量不相关情况下的功效分析

如方程5.14所示，协变量与干预指示变量不相关情况下的功效分析所需的先验信息更少。它仅包括协变量对因变量变异之解释的有效性（$\rho^2_{yx,w}$）。而且，由于干预组和协变量之间的相关假定为0，因而效应值不再需要修正，而是等于$\delta=\frac{\mu_1-\mu_0}{\sigma}$。[17]同样，方差膨胀也被消除了，所以方程5.12就约等于下述简化式：

$$\lambda \approx \underbrace{\delta}_{\text{效应值}} \underbrace{\sqrt{\frac{n}{2}}}_{\text{样本量}} \underbrace{\sqrt{\frac{1}{1-\rho^2_{yx,w}}}}_{\text{协方差效应}} \qquad [5.19]$$

概言之，协变量与干预指示变量不相关时，功效分析仅涉及效应值、样本量和协变量对因变量变异之解释的有效性。

求先验功效

此类情况下的功效计算与其他情况并无二致。本质上来说，一旦得到了非中心化参数，我们就利用合适自由度下的非中心化t分布，以确定第二类错误（β）所涵盖的区域面

积。一旦确定了这一点，那么很容易就得到功效为 $1-\beta$。

实例

假定我们的实例数据中干预指定变量与协变量不相关。实例数据分析中未修正的效应值约为 $\delta=0.678$。依然假定某个协变量与结果变量的相关为 $\rho_{yx,w}=0.983\ 3$，但与干预指示变量不相关，即 $\rho_{Tx}=0$。分析原始数据得到的非中心化参数为 2.397，与之相联系的自由度为 48，功效为 0.651。如果协变量与结果变量高度相关，达到 $\rho_{yx,w}=0.983\ 3$，那么新的非中心化参数为：

$$\lambda\approx\delta\sqrt{\frac{n}{2}}\sqrt{\frac{1}{1-\rho_{yx,w}^{2}}}\approx0.678\sqrt{\frac{25}{2}}\sqrt{\frac{1}{1-0.983\ 3^{2}}}\approx13.171$$

这将是我们的 t 统计量，与之相联系的功效接近 1。

求样本量

假定已有知识提供了有关协变量和结果变量之间相关的信息，那么我们可利用方程 3.9 对方程 5.19 进行变换，得到近似样本量的计算式：

$$n_z\approx\frac{2(z_{critical}-z_{\beta})^2}{\delta^2}(1-\rho_{yx,w}^{2}) \qquad [5.20]$$

然后就可以用上述计算式的结果来求更近似的样本量值：

$$n\approx\frac{2[t_{(2n_z-3)critical}-t_{(2n_z-3)\beta}]^2}{\delta^2}(1-\rho_{yx,w}^{2}) \qquad [5.21]$$

实例

假定我们期望效应值为 $\delta=0.75$,并相信我们所使用的某个协变量与干预变量不相关,但与结果变量相关,$\rho_{yx,w}=0.8$。若想在 $\alpha=0.01$ 水平上对我们的假设进行双尾检验,且想要获得 0.9 的功效(因此 $\beta=0.1$),我们首先根据方程 5.20 以及 $z_{0.995}$ 和 $z_{0.1}$ 值求近似样本量,得到每组约为 20 个个案。

$$n_z \approx \frac{2(z_{critical}-z_\beta)^2}{\delta^2}(1-\rho_{yx,w}^2)$$

$$n_z \approx \frac{2[2.576-(-1.282)]^2}{0.75^2}(1-0.8^2)=19.052$$

然后根据这个数值,利用分位数表,得到自由度为 35 所在行[18]的 t 分位数 $t_{(35)0.995}$ 和 $t_{(35)0.1}$,计算得到每组需包含约 21 个个案。

$$n \approx \frac{2[t_{(35)0.995}-t_{(35)0.1}]^2}{\delta^2}(1-\rho_{yx,w}^2)$$

$$n \approx \frac{2[2.724-(-1.306)]^2}{0.75^2}(1-0.8^2)\approx 20.788$$

求最低可检测效应

对方程 5.19 进行变换可得到最低可检测效应值计算式:

$$\delta_m \approx [t_{(2n-3)critical}-t_{(2n-3)\beta}]\sqrt{\frac{2}{n}}\sqrt{1-\rho_{yx,w}^2} \qquad [5.22]$$

实例

假定已知每组样本量为 $n=12$,且我们相信某个协变量

与干预变量不相关，但与结果变量相关，$\rho_{yx,w}=0.8$。若想在 $\alpha=0.01$ 水平上对我们的假设进行双尾检验，且想要获得 0.9 的功效（因此 $\beta=0.1$），我们就需要使用 $t_{(21)0.995}$ 和 $t_{(21)0.1}$。根据方程 5.22，我们可能检测到的最低效应为：

$$\delta_m \approx [2.813-(-1.323)]\sqrt{\frac{2}{12}}\sqrt{1-0.8^2} \approx 1.018$$

这意味着为了满足成功检验所需的标准，我们需要大于 1 个标准差的效应。

第5节 | 小结

上一章讨论了运用简单 t 检验来检验两组均值差异。本章则探讨了协变量对这一分析可能产生的影响,通过回归方法考察了协变量分析模型。我们发现协变量可能有所裨益,因为协变量会降低回归模型的残差方差,从而减小回归效应的抽样方差。然而,如果协变量与干预指示变量存在相关,那么多重共线性就会影响抽样方差。在这种情况下,方差膨胀因子就会受到干预指示变量和协变量的影响。随着这一相关的增加,方差膨胀因子会使回归效应的抽样方差增加,导致更低的统计功效。在实例数据中,我们看到这一协变量相关带来的益处远不如多重共线性的效应。

下一章将重新聚焦于组均值差异。然而,我们会面对更为复杂、多层的抽样设计。这一多层抽样设计也可以在不同的层次纳入协变量。

第6章

多层模型Ⅰ：二层聚类随机试验中的组均值差异检验

本章将引入聚类或多层设计。多层设计在实验研究中颇为常见,其原因可参见布卢姆(Bloom, 2005)的综述。多层设计在教育(O'Connell & McCoach, 2008)和健康(Donner & Klar, 2000; Murray, 1998)研究中尤为常见。虽然通常使用HLM(Raudenbuch, Bryk, & Congdon, 2004)这一类分层模型分析软件来拟合这类模型,但我们仍然用方差分析框架来讨论检验问题,以更好地理解此类模型功效分析的机制和参数估计。[19]本章的重点将放在二层设计上,但我们鼓励读者进一步阅读更多层设计的相关文献(如 Hedges & Rhoads, 2010)。本章的基本做法是为简单分析提供更多细节,而在之后的复杂分析中仅提供关键信息。

第1节 实例数据

聚类随机试验的实例分析来自附录中的一个小数据集。* 该数据乃"社会资本与儿童发展研究:一个随机控制试验"(Social Capital and Children's Development: A Randomized Controlled Trial)的一个小规模子样本(Gamoran, Turley, Turner, & Fish, 2012)。该研究从2008年至2013年在美国菲尼克斯和圣安东尼奥的52所学校开展调查。实例数据包括了八个随机抽取的凤凰城学校,在每个学校随机抽取五名有三年级数学成绩信息的学生。虽然实例数据仅包含了实际数据的很小一部分,但表中包含了"向公众开放"的学校和学生代码,因而这些研究可被重复。完整的公众开放数据文档可在ICPSR获取(研究编号为35481)。

该研究随机指定学校接受干预或按常规运行(因此这是一个聚类随机试验)。所谓实验干预是指接受一系列旨在增加一年级学生的父母和教师间社会资本的措施。具体而言,干预措施即"家校一体"(Families and Schools Together, FAST)项目。实例中的结果变量数据是被观测学生在三年级时的数学成绩。[20]

---

* 译者把 trial 翻译为"试验",以区别于 experiment(实验)。一般而言,trial(试验)指涉及人的 experiment(实验)。——译者注

第 2 节 | 以方差分析来理解单层检验

多层(如混合或分层)模型可追溯其源头到多因素方差分析。因此,为了给多层模型的讨论做好充分准备,我们首先以方差分析视角来简要回顾第 4 章中的组均值差异检验。

简单随机样本的单因素方差分析 *F* 检验

检验组均值差异的另一种方法是方差分析。很多研究者偏爱用方差分析来模型化干预效应,因为它可使研究从两组比较拓展到更多有趣的实验场景。

回顾一下方程 4.1,其中每组的干预效应标记为 $\tau_j = \mu_j - \mu$,观测值与组均值的差异为 $e_{ij} = y_{ij} - \mu_j$。因此可以把数据生成过程用下式表达:

$$y_{ij} = \mu + (\mu_j - \mu) + (y_{ij} - \mu_j) \qquad [6.1]$$

这一式子让我们可以把每个观测值相对于总估计均值的离差平方和(即总平方和 SST)分解为两部分:一是每组估计均值相对于总估计均值的离差平方和(以组样本量进行加权,即组间平方和 SSB),二是每个观测值相对于每组估计均值的离差平方和(即组内平方和 SSW)。

$$\underbrace{\sum_j \sum_i (y_{ij}-\bar{y})^2}_{\text{SST}} = \underbrace{n\sum_j (\bar{y}_j-\bar{y})^2}_{\text{SSB}} + \underbrace{\sum_j \sum_i (y_{ij}-\bar{y}_j)^2}_{\text{SSW}} \qquad [6.2]$$

单因素方程分析所用的 F 检验是干预效应的估计方差与总体方差估计值(或组内方差)的比值。F 比值的分子，即干预效应方差，被定义为组间平方和的均值(即组间均方 MSB)：

$$\text{MSB}=\frac{n\sum_j \tau_j^2}{p-1} \qquad [6.3]$$

且限制条件为 $\sum_j \tau_j=0$。已知 τ_j 是总均值与组均值之差，因此方程 6.3 可写作：

$$\text{MSB}=\frac{n\sum_j (\bar{y}_j-\bar{y})^2}{p-1}$$

F 比值的分母，即组内方差，被定义为组内平方和的均值(即组内均方 MSW)：

$$\text{MSW}=\frac{\sum_j \sum_i (y_{ij}-\bar{y}_j)^2}{pn-p} \qquad [6.4]$$

对这一表达式我们颇为熟悉，与估计总体方差的方程 4.5 非常类似。因此 F 比值被定义为：

$$F=\frac{\text{MSB}}{\text{MSW}} \qquad [6.5]$$

这一检验统计量的分子自由度为 $p-1$，分母自由度为 $pn-p$。

这些估计值通常都集中呈现在类似表 6.1 的方差分析表中(参见 Casella, 2008)。该表清晰展现了方差分析流程的

不同部分,并把总平方和分解成组间平方和与组内平方和。我们可以使用表 6.2 中的数字来进行 F 检验的计算。在实例数据中,组间差异为 2.35(参见表 6.7),n 为 20,故而 MSB 为 $\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2=\frac{n}{2}2.35^2=55.225$。这是因为[21]:

$$\mathrm{MSB}=\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2 \qquad [6.6]$$

MSW 为 4.704,是表 6.7 中的均方误差根 2.169 的平方。因此得到 F 比值等于 11.740。由于自由度为(1, 38),所以得到 $p<0.001$,这与 t 检验的 p 值相等。

表 6.1 单因素方差分析表

来源	自由度 df	平方和	均方	F 检验
组间	$p-1$	$\mathrm{SSB}=n\sum_j(\bar{y}_j-\bar{y})^2$ $=n\sum_j\tau_j^2$	$\mathrm{MSB}=\frac{\mathrm{SSB}}{p-1}$	$F=\frac{\mathrm{MSB}}{\mathrm{MSW}}$
组内	$np-p$	$\mathrm{SSW}=\sum_j\sum_i(y_{ij}-\bar{y}_j)^2$	$\mathrm{MSW}=\frac{\mathrm{SSW}}{np-p}$	
总计	$np-1$	$\mathrm{SST}=\sum_j\sum_i(y_{ij}-\bar{y})^2$		

表 6.2 干预后三年级数学成绩的单因素方差分析表

	自由度 df	平方和	均方	F	$\Pr(>F)$
“家校一体”项目	1.000	55.225	55.225	11.740	0.001
残差	38.000	178.750	4.704		

由此,我们就很容易得知 F 比值是 t 比值的平方。可以把 t 检验表述为 $t=\frac{\bar{y}_1-\bar{y}_0}{\hat{\sigma}\sqrt{\frac{2}{n}}}$,对其进行平方并稍作变换,就可

得到 $t^2=\frac{\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2}{\hat{\sigma}^2}$。该式即单因素方差分析的 F 比值。这一练习使我们知道,当仅有两个实验组时,把 F 统计量取根号就可以转化为 t 统计量,恰如第2章所示。

在讨论聚类试验之前,我们花点时间探讨一下期望均方。在单因素固定方差分析模型中,我们仅有一个方差来源,即总体方差 σ^2。而在多层模型中,我们要处理包含多个方差来源的情况。

期望均方

上述分析背后的理念是以数据来检验组间方差(即干预效应)为0的零假设。F 检验是方差间的比值,是期望均方之间比值的一部分。这里不进行均方的公式推导,我们鼓励读者去自行查阅实验设计相关著作,其中一本优秀著作是柯克的研究(Kirk, 1995)。譬如,表6.3呈现了单因素固定方差分析的期望均方。

表 6.3　单因素方差分析的期望均方表

来源	期望均方	期望 F 检验
组间	$E(\text{MSB})=\sigma_e^2+n\frac{\sum_j\tau_j^2}{p-1}$	$F=\frac{\sigma_e^2+n\frac{\sum_j\tau_j^2}{p-1}}{\sigma_e^2}$
组内	$E(\text{MSW})=\sigma_e^2$	

单因素方差分析(假定组别是固定的,而非随机选取自一个更大的组别库)中 $p=2$ 的组间期望均方[标记为 $E(\text{MSB})$],本质上是组内方差加上引入干预效应导致的方差:

$$E(\text{MSB})=\sigma_e^2+n\frac{\sum_j\tau_j^2}{p-1}=\sigma_e^2+\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2 \quad [6.7]$$

组内期望均方即组内方差或总体方差:

$$E(\text{MSW})=\sigma_e^2 \quad [6.8]$$

因此,期望的检验值就等于1加上实际的检验值:

$$\frac{E(\text{MSB})}{E(\text{MSW})}=\frac{\sigma_e^2+\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2}{\sigma_e^2}=1+\frac{\frac{n}{2}(\bar{y}_1-\bar{y}_0)^2}{\sigma_e^2} \quad [6.9]$$

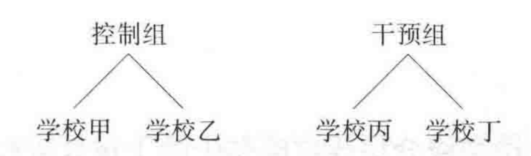
若干预引入的方差为0,该表达式的期望值就是1的零假设比值。所以,该检验就是1加上比值$\frac{n(\bar{y}_1-\bar{y}_0)^2}{2\sigma_e^2}$。在更为复杂的模型中,我们使用期望均方,以更好地理解这些检验。

如第2章所示,对这一表达式取平方根就能得到自由度为$2n-2$的t检验。这是因为分子自由度为1(因为$p-1=2-1=1$,参见表6.1)的F检验,等同于自由度为F检验之分母自由度($pn-p=2n-2$)的t检验的平方。当把自由度为(1, df_d)的F检验转化为t检验时,我们总是把分母的自由度df_d用到t检验中。

第3节 聚类随机试验的多层混合模型

在这一部分,我们讨论聚类随机试验。首先说明数据生成过程,然后详述这类分析通常所使用的多层回归模型。接下来利用双因素混合方差分析模型来计算功效的参数。

之所以考虑这类"分层"线性模型,是因为在固定实验组中存在完全的聚类嵌套。譬如,有两个实验组,且每个实验组包含两个聚类单元,那就需要根据嵌套于实验中的随机分组来思考这一数据结构。为了进一步具体说明,以图6.1为例,实验组包含了"控制"和"干预"两种情况,而四个学校则嵌套于不同实验组中。



注:其中实验组包括控制组和干预组,聚类单元为学校。

图6.1　嵌套于实验组的多层聚类图示

假定共有 $j=\{1, 2, \cdots, p\}$ 个实验组(本书中 $p=2$),每组包含 $k=\{1, 2, \cdots, m\}$ 个聚类单元(cluster),且每个聚类单元包含 $i=\{1, 2, \cdots, n\}$ 单位(unit)。这一模型的数据生

成过程就是在方程 4.1 中增加一个组成项和下标:

$$y_{ik(j)}=\mu+\tau_j+b_{k(j)}+e_{ik(j)} \tag{6.10}$$

其中 $y_{ik(j)}$ 是指派给实验组 j 的聚类单元 k 的第 i 个单位,$\tau_j=\mu_j-\mu$ 是组 j 的干预效应(和前面一样),$b_{k(j)}=\mu_{k(j)}-\mu_j$ 是控制干预效应后聚类 k 的随机效应,而 $e_{ik(j)}=y_{ik(j)}-\mu_{k(j)}$ 是组内的聚类单元内的残差项。把 $b_{k(j)}$ 标记为随机效应是因为我们假定聚类单元是由某个随机程序选定的,所以使用了一个罗马字母,而非希腊字母。然而,我们仍然假定分析的是所有可能的实验组,因此 τ_j 是一个固定效应,以希腊字母标记。我们应当搞清楚称这些模型为"混合"模型的原因。因为除了误差项,这一模型现在还"混合"了固定效应和随机效应。

由于我们的数据是通过随机程序观测到的,即随机选取聚类单元以及在聚类单元内部随机选取单位,所以除了干预效应之外还存在两个方差来源。这些方差组成部分乃基于随机效应的方差。具体而言,即 $e_{ik(j)}\sim N(0,\ \sigma_e)$ 和 $b_{k(j)}\sim N(0,\ \sigma_b)$。因此,$\sigma_e^2$ 和 σ_b^2 是方差的组成部分,除干预效应外,结果变量的总方差为 $\sigma^2=\sigma_b^2+\sigma_e^2$。

线性混合模型或分层线性模型中的干预效应检验

聚类随机试验可以用双因素方差分析模型进行分析,但如今这类分析通常使用线性混合模型(LMM; McCulloch & Searle, 2001)或分层线性模型(HLM; Raudenbush & Bryk, 2002),这些模型使用最大似然法或限制性最大似然法的迭

代技术，估计方法的具体细节已超出本书讨论范围，不作赘述。

线性混合模型和分层线性模型是等价的两个模型，只是使用了不同的符号体系。在线性混合模型的符号体系中，我们在方程 4.7 那样的线性回归中增加下标和随机效应：

$$y_{ik(j)}=\gamma_0+\gamma_1 T_{k(j)}+b_{k(j)}+e_{ik(j)} \qquad [6.11]$$

在分层线性模型的符号体系中，我们设想以不同的关联模型来代表每“层”的数据。就聚类单元中的单位（譬如学校中的学生）而言，我们把层 1 模型设定为[22]：

$$y_{ik(j)}=\pi_{0k(j)}+e_{ik(j)} \qquad [6.12]$$

其中，$\pi_{0k(j)}$ 为实验组 j 中的聚类 k 所包含单位的结果变量平均值，而 $e_{ik(j)}$ 为组—聚类内部的正态残差分布，其标准差为 σ_e。

然后假定每个聚类组的均值是一组协变量（即干预指示变量）和一个随机效应的函数：

$$\pi_{0k(j)}=\gamma_{00}+\gamma_{01} T_{k(j)}+b_{0k(j)} \qquad [6.13]$$

其中，γ_{00} 是控制组中各聚类（$T=0$）均值的总平均值，γ_{01} 是干预组各聚类均值之平均值与控制组各聚类均值之平均值的差异。最后，$b_{0k(j)}=\pi_{0k(j)}-[\gamma_{00}+\gamma_{01} T_{k(j)}]$，即组中各聚类均值与干预组或控制组均值的差异。倘若把方程 6.13 代入方程 6.12 以替代 $\pi_{0k(j)}$，便得到一个与方程 6.11 等价的方程。

随机效应的方差是估计方差的组成部分。随机效应 $e_{ik(j)}$ 服从均值为 0、标准差为 σ_e 的正态分布，随机效应 $b_{0k(j)}$ 服从均值为 0、标准差为 σ_b 的正态分布。

最大似然方法使得功效分析颇为困难。然而,大多数软件在运算线性混合模型和分层线性模型时所使用的限制性最大似然法,却能进行类似于方差分析的估计(Raudenbush & Bryk, 2002)。因此,方差分析的期望均方在设计研究时颇有裨益。

表 6.4 双因素混合方差分析的期望均方表(干预来源固定但聚类来源随机)

来　源	期望均方
组　间	$E(MS_\tau)=\sigma_e^2+n\sigma_b^2+nm\dfrac{\sum_j \tau_j^2}{p-1}$
聚类之间	$E(MS_b)=\sigma_e^2+n\sigma_b^2$
聚类内部	$E(\mathrm{MSW})=\sigma_e^2$

聚类随机试验中干预效应的混合双因素方差分析检验

单因素方差分析把某个结果变量的方差沿着单个因素进行分解。双因素方差分析则沿着两个因素对某个结果变量的方差进行分解。在此种情况下,其中一个因素是聚类单元(即所有聚类单元的一个随机样本),另一个因素是干预组(即所有可能干预组的一个固定集合)。在讨论平方和及其计算细节之前,我们先回到期望均方,它可以由样本规模和方差成分进行定义。

对于混合分层检验,我们用 MS_τ 与 MS_b 的比值来检验干预效应,故而期望的检验为$\dfrac{E(MS_\tau)}{E(MS_b)}$,因为分母 MS_b 的组成部分可用于分离分子 MS_τ 中的干预效应(Schultz, 1955)。期望均方的定义如表 6.4 所示(Kirk, 1995)。当 $p=2$ 时,

MS_τ 的期望均方为：

$$E(MS_\tau) = \sigma_e^2 + n\sigma_b^2 + nm\frac{\sum_j \tau_j^2}{p-1} = \sigma_e^2 + n\sigma_b^2 + \frac{n}{2}m(\mu_1 - \mu_0)^2 \quad [6.14]$$

而聚类单元间的期望均方为：

$$E(MS_b) = \sigma_e^2 + n\sigma_b^2 \quad [6.15]$$

因此,检验统计量为下述比值：

$$F = \frac{\frac{1}{2}nm(\mu_1 - \mu_0)^2}{\sigma_e^2 + n\sigma_b^2} \quad [6.16]$$

聚类随机试验中干预效应的 t 检验

F 检验(方程 6.16)进行变换并取平方根,就变成 t 统计量的形式：

$$t = \frac{\mu_1 - \mu_0}{\sqrt{\frac{2}{nm}(\sigma_e^2 + n\sigma_b^2)}} \quad [6.17]$$

其中,干预组差异的抽样方差为：

$$\hat{\text{var}}\{\bar{y}_1 - \bar{y}_0\} = \frac{2}{nm}(\hat{\sigma}_e^2 + n\,\hat{\sigma}_b^2) \quad [6.18]$$

这一 t 检验的自由度为 $2m-n$。这里使用聚类单元的数量 $2m$,而非总样本量 $2nm$。使用聚类单元数量是因为在 F 统计量中,我们除的是聚类单元间的均方(MS_b),而非组一聚类单元内部的均方。

第4节 | 聚类随机试验的功效参数

遗憾的是,方程 6.17 并非进行功效分析的理想公式,因为其中未包含无标尺参数。首先,需要一个效应值。一个可行办法是像方程 3.10 那样,用均值差除以总的标准差(Hedges, 2007):

$$\delta=\frac{\mu_1-\mu_0}{\sigma}=\frac{\mu_1-\mu_0}{\sqrt{\sigma_b^2+\sigma_e^2}} \qquad [6.19]$$

其次,利用方差组成部分,我们可以构建另一个无标尺参数。聚类单元之间的方差 σ_b^2 是聚类单元内各单位的组内协方差(McCulloch & Searle, 2001)。它表示聚类单元内部各单位相互间的相关程度。我们可以用总变异对这一度量进行标准化,得到一个实际的相关系数。这一无标尺参数即组内相关系数 ρ_{intra}:

$$\rho_{intra}=\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2} \qquad [6.20]$$

这一相关系数使得聚类随机试验的功效分析成为可能。组内相关系数的含义乃是代表了聚类单元内部各单位之间的相关性。

若假定总方差 $\sigma^2=1$,那就能简单地用均值差作为效应

值。而且,有了组内相关(方程 6.20),就能把方差组成部分转化成无标尺参数 $\sigma_b^2=\rho_{intra}$，$\sigma_e^2=1-\rho_{intra}$。* 据此,就可以把 t 检验进行如下表达:

$$t=\delta\sqrt{\frac{m}{2}}\sqrt{\frac{1}{[(1-\rho_{intra})/n]+\rho_{intra}}}=\delta\sqrt{\frac{nm}{2}}\sqrt{\frac{1}{1+(n-1)\rho_{intra}}}$$

其中,$1+(n-1)\rho_{intra}$ 为"设计效应"(Kish, 1965)。

一般而言,某抽样设计的"设计效应"是一个比值,即利用恰当方法得到某个估计值之抽样方差,与假定来自简单随机样本分析的估计值之抽样方差的比值。在这里,设计效应可以通过方程 6.18 与方程 4.18 的比值得到,前提是假定聚类设计中的 mn 等于简单随机样本中的 n,且 $\sigma_b^2+\sigma_e^2=\sigma^2$。

聚类随机试验的非中心化参数等于:

$$\lambda=\underbrace{\delta}_{\text{效应值}}\underbrace{\sqrt{\frac{nm}{2}}}_{\text{样本量}}\underbrace{\sqrt{\frac{1}{1+(n-1)\rho_{intra}}}}_{\text{设计效应}} \qquad [6.21]$$

其中,第一项 δ 是效应值;第二项是样本量,n 是每个聚类单元所包含的单位数,m 是每个实验组所包含的聚类单元数;第三项是设计效应,ρ_{intra} 是组内相关。

聚类随机试验中使用不相关的协变量

如第 4 章单层模型所示,使用与干预指示变量不相关的控制变量能改善功效,因为减小了条件方差。在聚类随机试验中也是如此。假定使用了一个协变量 x,与随机化的干预

---

* 因为假定了 $\sigma^2=\sigma_b^2+\sigma_e^2=1$。——译者注

分组无关。为了最大化聚类随机试验中该协变量的益处，在回归模型中纳入聚类均值对中的单位层面的 x 值及其聚类均值 $\bar{x}_{k(j)}$：

$$y_{ik(j)}=\gamma_0+\gamma_1 T_{k(j)}+\gamma_2[x_{ik(j)}-\bar{x}_{k(j)}]+\gamma_3\bar{x}_{k(j)}+b^*_{k(j)}+e^*_{ik(j)} \quad [6.22]$$

我用星号（*）来提醒读者，由于纳入了协变量，随机效应有所缩减。由于（干预组内）协变量与随机效应间存在相关，因此方差的组成相应地有所缩减。譬如，$b^*_{k(j)}$ 的方差为：

$$\sigma_b^{2*}=\sigma_b^2(1-R^2_{cluster})$$

而 $e^*_{ik(j)}$ 的方差为：

$$\sigma_e^{2*}=\sigma_e^2(1-R^2_{unit})$$

其中，$R^2_{cluster}$ 是聚类单元层面的方差被聚类均值协变量所解释的比例，而 R^2_{unit} 是被聚类内部的协变量所解释的方差比例。

因此，差值的方差（方程 6.18）就变为：

$$\hat{\text{var}}\{\bar{y}_1-\bar{y}_0\}=\frac{2}{nm}[\hat{\sigma}_e^2(1-R^2_{unit}]+n\hat{\sigma}_b^2[1-R^2_{cluster})]$$

我们还是可以使用无标尺参数把上式转化成更易处理的形式。若进行下述替换，$\sigma_b^2=\rho_{intra}$，$\sigma_e^2=1-\rho_{intra}$，则差值的方差就可以变换成下式：

$$\hat{\text{var}}\{\delta\}=\frac{2}{nm}\{1+(n-1)\rho_{intra}-[R^2_{unit}+(nR^2_{cluster}-R^2_{unit})\rho_{intra}]\} \quad [6.23]$$

从而得到非中心化参数为(Hedges & Rhoads，2010)：

$$\lambda=\delta\sqrt{\frac{nm/2}{1+(n-1)\rho_{intra}-[R_{unit}^{2}+(nR_{cluster}^{2}-R_{unit}^{2})\rho_{intra}]}}$$

[6.24]

该检验的自由度为 $2m-2-q$，其中 q 为模型中所使用的聚类层面协变量的数量。

第5节 | 聚类随机试验的实例分析

在这一节,我们回到本章的实例数据,通过分析更好地理解这些参数。样本中的干预组和控制组分别包括 $m=5$ 个学校,每个学校包含 $n=4$ 个学生。

表6.5提供了结果变量的基本描述统计。就均值而言,我们看到干预组在数学测评分数上比控制组高2.35分。这一结果在表6.7中得到验证,参与"家校一体"项目的系数为2.35。在一般最小二乘(OLS)模型中,这一结果的 t 检验值为 $2.350/0.686=3.426$,等于表6.2中所报告的 F 检验值的平方根。

表6.5 干预后三年级数学成绩的概要统计

	N	均值	标准差
未参与项目(控制组)	20.000	6.100	2.532
"家校一体"项目(干预组)	20.000	8.450	1.731
总样本	40.000	7.275	2.449

表6.7的"混合1"模型提供了混合模型的估计值检验,得到 t 值为 $2.350/0.762=3.084$,小于OLS模型的 t 值,显著性没那么强了。原因在于差值的方差增加了 $0.762^2/0.686^2=1.234$ 倍。这是设计效应,可根据方差组成要素来计算组内的相关系数得到。学校层面的方差是0.351,而学校内部的

残差方差为 4.408,所以组内相关系数为 $\rho_{intra}=\frac{0.351}{0.351+4.408}=0.074$。根据组内相关系数,就可以计算设计效应,$1+(n-1)\rho_{intra}=1+(4-1)0.074=1.222$(与 1.234 的差异源自四舍五入)。

接下来,我们考虑一个协变量,其概要统计参见表 6.6。从中可见,干预组和控制组的前测均值是相等的。因此,前测与干预指示变量不相关。在表 6.7 的"混合 2"模型中,"家校一体"项目的系数没有发生变化。这也是可以预见的,因为协变量不相关。然而,其标准误要小于其他模型。原因在于,协变量减小了学校和学校内部(残差)层面的方差组成部分。在学校层面,聚类协变量减少了 57.8%的学校间方差,$R^2_{cluster}=1-(0.148/0.351)=0.578$;单位层面协变量减少了 26.0%的学校内部方差,$R^2_{unit}=1-(3.263/4.408)=0.260$。因此,设计效应为:

$$1+(n-1)\rho_{intra}-[R^2_{unit}+(nR^2_{cluster}-R^2_{unit})\rho_{intra}]$$

代入计算得到 $1+(4-1)0.074-[0.260+(4\times0.578-0.260)0.074]=0.810$。这一结果可通过模型估计方差的比值得到确认,$0.621^2/0.686^2=0.819$(与 0.810 的差异源自四舍五入)。

表 6.6 干预前一年级数学成绩的概要统计

	N	均值	标准差
未参与项目(控制组)	20.000	3.500	1.235
"家校一体"项目(干预组)	20.000	3.500	1.318
总样本	40.000	3.500	1.261

表 6.7　预测三年级数学成绩的模型

	OLS	混合 1	混合 2
截距	6.100***	6.100***	5.956**
	(0.485)	(0.539)	(1.873)
"家校一体"项目对比控制组	2.350**	2.350**	2.350***
	(0.686)	(0.762)	(0.621)
一年级数学能力			0.946***
			(0.279)
一年级能力的学校均值			0.041
			(0.520)
R^2	0.236		
N	40	40	40
MSE 的平方根	2.169		
聚类数量		10	10
σ_b^2		0.351	0.148
σ_e^2		4.408	3.263

注:一年级数学成绩进行了组均值对中。括号中为标准误。* $p<0.05$, ** $p<0.01$, *** $p<0.001$。

第 6 节 | 聚类随机试验的功效分析

求先验功效

一旦计算出了非中心化参数（方程 6.24），接下来的功效分析就和本书其他部分的 t 检验相同了。一旦得到了 λ，检验的功效就可以通过计算机用第 4 章所介绍的 H 函数进行计算，即方程 4.23 和方程 4.24。需要注意的是，这一检验的自由度是 $2m-2-q$，其中 q 是检验中所用到的聚类层面协变量的数量。若未使用协变量，就可以把 q、$R^2_{cluster}$ 和 R^2_{unit} 简单设置为 0。

求样本量

在两层模型中有两个样本量：聚类的数量和聚类中单位的数量。如下文所述，增加单位数量对功效的助益有限，而增加聚类的数量则收益颇大。这一部分内容聚焦于，聚类中单位数量固定时，如何计算聚类的数量。

增加单位还是聚类？

设计聚类随机试验时经常遇到的一个问题是需要权衡

增加单位数量还是增加聚类数量。这并不是一个很直观的问题，主流的建议是增加聚类的数量，而非聚类单元中的单位数量。[23]这样做的原因可参考图 6.2。在该图中，我们把效应值和组内相关系数设为固定值，$\delta=0.3$，$\rho_{intra}=0.2$，然后画出不同 m 和 n 组合情况下的功效曲线。

在图 6.2 中，我们画出了功效水平 0.4 到 0.9 之间的几条线。功效曲线 0.4 以左的区域代表 m 和 n 组合得到的功效小于 0.4。0.4 和 0.5 之间的区域代表 m 和 n 组合得到的功效在 0.4 到 0.5 之间。譬如，若某设计的每个实验组包含 20 个聚类单元，每个聚类包含 15 个单位（因此总样本量为 $2\times20\times15=600$），该设计的功效就小于 0.5，但大于 0.4。

仔细研究图 6.2，就会发现增加聚类中的单位数量会对功效产生影响，但仅在单位数量达到大约 30 之前。换言之，单位数量达到 30 之后，如果我们纵向来看功效曲线，它并未发生多大变化。然而，如果横向来看，功效曲线在横轴的大部分区间都有明显变化。这就意味着增加实验组的聚类数量比增加聚类中单位的数量更有用。

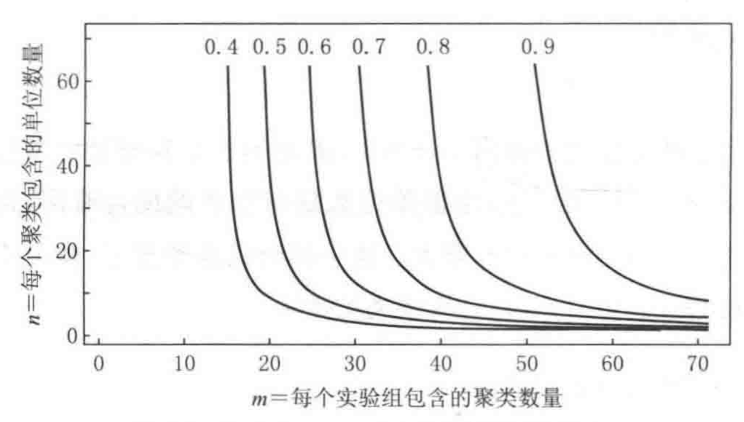


图 6.2 $\delta=0.3$、$\rho_{intra}=0.2$ 的聚类随机试验中，m 和 n 不同取值情况下的功效曲线图

求单位规模固定时的聚类数量

与本书其他的 t 检验一样，我们首先求标准正态分布近似值。对方程 6.24 进行变换，得到聚类中单位数量固定情况下求聚类数量的公式：

$$m_Z \approx \frac{2(z_{critical}-z_{\beta})^2 D}{n\delta^2} \quad [6.25]$$

其中 D 为设计效应：

$$D=1+(n-1)\rho_{intra}-[R^2_{unit}+(nR^2_{cluster}-R^2_{unit})\rho_{intra}] \quad [6.26]$$

有了正态分布近似值之后，就可以用 t 分布的分位数来改进估计值：

$$m \approx \frac{2[t_{(2m_Z-2-q)critical}-t_{(2m_Z-2-q)\beta}]^2 D}{n\delta^2} \quad [6.27]$$

其中 D 为设计效应（同方程 6.26），q 是协变量的数量（用于得到 t 分布分位数）。

实例

倘若我们要为一门新课设计聚类随机试验，这将会是一个双尾检验，$\alpha=0.05$，希望达到的功效为 0.8。我们计划把学校随机分到各实验组中，每个学校有 $n=30$ 个学生参与。我们期望实验组的数学成绩会有 $\delta=0.5$ 个总体标准差的改善。我们还期望学校内部的数学成绩组内相关系数约为 $\rho_{intra}=0.2$，并且学生（单位）层面的协变量约解释学生层面方差比例为 $R^2_{unit}=0.25$，解释学校（聚类）层面方差比例为 $R^2_{cluster}=$

0.64。第一步是计算设计效应 D：

$$D=1+(n-1)\rho_{intra}-[R_{unit}^2+(nR_{cluster}^2-R_{unit}^2)\rho_{intra}]$$

$$D=1+(30-1)0.2-[0.25+(30\times0.64-0.25)0.2]=2.76$$

接下来,我们利用正态分布近似值求出所需聚类数量的近似值。这意味着 $z_{critical}=z_{0.975}=1.96$ 和 $z_{0.2}=-0.842$。基于这些求出来的值和设计效应,我们就可以计算:

$$m_Z\approx\frac{2(z_{critical}-z_\beta)^2D}{n\delta^2}\approx\frac{2[1.96-(-0.842)]^2 2.76}{30\times0.5^2}\approx5.778$$

即每个实验组需要包含六所学校(聚类单元)。然后利用这个值和附录中的 t 分布分位数来修正我们的估计,自由度为 $2m_z-2-q=2\times6-2-1=9$($q=1$ 表示聚类层面有一个协变量)。因此,$t_{(2m_z-2-q)critical}=t_{(9)0.975}=2.262$,$t_{(2m_z-2-q)\beta}=t_{(9)0.2}=-0.883$,而且,

$$m\approx\frac{2[t_{(2m_z-2-q)critical}-t_{(2m_z-2-q)\beta}]^2D}{n\delta^2}$$

$$\approx\frac{2[2.262-(-0.883)]^2 2.76}{30\times0.5^2}\approx7.280$$

即每个实验组需包含八所学校(聚类单元),所以共需要 16 所学校。这一研究设计的实际功效为 0.861,要大于 0.8,这是因为我们把样本量向上取整了。

求最低可检测效应

和前面一样,我们对方程 6.24 进行变换,得到求最低可检测效应(MDES)的公式:

$$\delta_m = [t_{(2m-2-q)critical} - t_{(2m-2-q)\beta}]\sqrt{\frac{2D}{nm}} \qquad [6.28]$$

其中 D 为上文所定义的设计效应（方程 6.26），q 为协变量的数量（用于得到 t 分布分位数）。

实例

倘若我们的资源仅允许每个实验组纳入七所学校。已知我们在计算步骤中通过取整来求得样本量，那么在总样本量为 14 所学校的情况下，我们可以计算得到一个可接受的最低可检测效应。利用和前面一样的设计效应（$D=2.76$），然后把 t 分布分位数调整为 $t_{(2m_z-2-q)critical}=t_{(11)0.975}=2.201$，$t_{(2m_z-2-q)\beta}=t_{(11)0.2}=-0.876$。我们还是计划每所学校纳入 $n=30$ 个学生。则最低可检测效应值为：

$$\delta_m = [t_{(2m-2-q)critical} - t_{(2m-2-q)\beta}]\sqrt{\frac{2D}{nm}}$$

$$\delta_m = [2.201-(-0.876)]\sqrt{\frac{2\times 2.76}{30\times 7}} = 0.499$$

可见，效应值接近目前这个例子中所要求的 0.5。

第 7 节 | 小结

本章探讨了多层设计,尤其是聚类单元(层 2)进行随机化的两层设计。虽然这类研究通常使用混合回归程序进行分析,但利用方差分析框架能使多层设计的功效计算更容易理解。因此,本章包含方差分析和期望均方的简要回顾。利用期望均方,本章计算了聚类随机试验的非中心化参数。聚类随机试验设计的一个关键参数是聚类单元内部的单位之间的相关程度,这被操作化为组内相关系数。组内相关系数越高,设计效应越大,从而相应地导致抽样方差增加。另一个需要考虑的结果是,一般来说,增加聚类数量比增加聚类内部的单位数量更有收益。一如既往,纳入与干预指示变量不相关的协变量有利于降低必要的样本规模。



多层模型Ⅱ：二层多点随机试验中的组均值差异检验

不同于第 6 章对聚类单元进行随机化,二层设计的另一个选择是对聚类内部的单位进行随机化。这实际上意味着每个聚类(或“研究点”)就是一个小型试验。* 这使得研究者可以估计干预效应如何在不同研究点之间变化。所谓的多点随机试验(multisite randomized trials)在社会实验和健康实验中也非常流行(Raudenbush & Liu, 2000)。本章将讨论这一类设计的功效分析。

假定每个聚类包含 $j=\{1, 2, \cdots, p\}$ 个实验组(本书中 $p=2$),每个实验组包含 $i=\{1, 2, \cdots, n\}$ 个单位,共有 $k=\{1, 2, \cdots, m\}$ 个聚类(研究点)。该模型的数据生成过程不同于方程 6.10,增加了一个交互项,而且下标也有所不同:

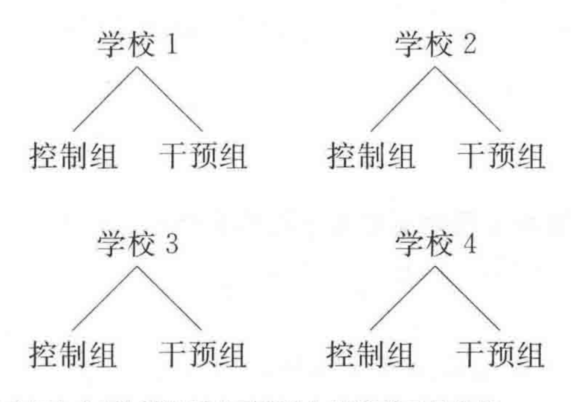
$$y_{ijk}=\mu+\tau_j+b_k+(\tau b)_{jk}+e_{ijk} \qquad [7.1]$$

其中,y_{ijk} 是聚类单元 k 中实验组 j 的第 i 个单位的结果变量取值,$\tau_j=\mu_j-\mu$ 是实验组 j 的干预效应(与前文相同),$b_k=\mu_k-\mu$ 是聚类 k 的随机效应,$(\tau b)_{jk}=\mu_{jk}-\mu_j-\mu_k+\mu$ 是干预效应在不同聚类中的随机效应,$e_{ijk}=y_{ijk}-\mu_{jk}$ 是聚类内的组内残差项。之所以把 b_k 和交互项 $(\tau b)_{jk}$ 作为随机效应,是

* 多点随机试验中也把聚类称为“研究点”(site)。——译者注

因为我们假定聚类是通过某种随机程序选取的。然而,我们还假定了所分析的是所有可能的实验组,因此 τ_j 本身是一个固定效应。

图7.1是对这一设计的图形化展示,其中每个聚类都包含一个干预效应。如图7.1所示,共有四所学校,每所学校包含一个干预组和一个控制组。这一模型包含了三个方差组成要素。首先是 $e_{ijk} \sim N(0, \sigma_e)$ 和 $b_k \sim N(0, \sigma_b)$,其次是 $(\tau b)_{jk} \sim N(0, \sigma_{\tau b})$。*



注:实验组包含"控制组"和"干预组",聚类单元是学校。

图7.1 实验组嵌套于聚类中的多点试验图示

线性混合模型或多层线性模型的干预效应检验

犹如聚类随机试验那样,我们可以估计一个混合模型进行分析,以检验干预效应。使用混合模型标记的回归模型如下所示。聚类单元 k 中实验组 j 的第 i 个单位的结果变量取值 y_{ijk},被建构为包含干预指示变量 T、聚类随机效应 b_k、干

* 原书中组内残差 e_{ijk} 方差的下标有误,已修改。——译者注

预效应的随机效应$(Tb)_{jk}$和组内残差e_{ijk}的方程：

$$y_{ijk}=\gamma_0+\gamma_1 T_{ijk}+b_k+(Tb)_{jk}+e_{ijk} \qquad [7.2]$$

分层线性模型框架能更好地理解交互项具体如何运行及其含义。运用分层线性模型框架，在“层 1”模型中，结果变量是一个包含不同聚类控制组均值π_{0k}、不同聚类干预效应π_{1k}和聚类内部残差e_{ijk}的方程：

$$y_{ijk}=\pi_{0k}+\pi_{1k}T_{ijk}+e_{ijk} \qquad [7.3]$$

然后再设置“层 2”模型。特定聚类的控制组均值等于聚类控制组均值的平均值加上聚类随机效应的方程：

$$\pi_{0k}=\gamma_{00}+b_{0k} \qquad [7.4]$$

类似地，聚类干预效应是关于聚类干预效应均值的方程：

$$\pi_{1k}=\gamma_{10}+b_{1k} \qquad [7.5]$$

方程 7.5 非常重要，因为它表明随机效应b_{1k}为何是聚类干预效应和聚类干预效应均值的差值。当我们用方程 7.4 代替方程 7.3 中的π_{0k}，用方程 7.5 代替方程 7.3 中的π_{1k}，就得到了下述混合模型：

$$y_{ijk}=\gamma_{00}+b_{0k}+(\gamma_{10}+b_{1k})T_{ijk}+e_{ijk}$$
$$y_{ijk}=\gamma_{00}+b_{0k}+\gamma_{10}T_{ijk}+b_{1k}T_{ijk}+e_{ijk}$$

其中，下述符号标记与混合模型中的标记（即方程 7.2）是对等的：$\gamma_{00}=\gamma_0$，$\gamma_{10}=\gamma_1$，$b_{1k}T_{ijk}=(Tb)_{jk}$。

随机效应的方差是方差组成部分的估计值。随机效应$e_{ik(j)}$服从均值为 0、标准差为σ_e的正态分布，随机效应b_{0k}服从均值为 0、标准差为σ_b的正态分布，随机效应$(Tb)_{jk}$服从

均值为0、标准差为$\sigma_{\tau b}$的正态分布。

和聚类随机试验一样，上述这些模型可以用限制性最大似然法进行估计。同样，类似于聚类随机试验，当考虑等价的方差分析模型时，功效分析会变得更容易。

多点随机试验中干预效应的混合双因素方法分析检验

表7.1呈现了该模型的期望均方。主干预效应的期望F检验是主效应期望均方与交互项期望均方的比值(Kirk，1995)：

$$\frac{E(MS_{\tau})}{E(MS_{\tau b})}=\frac{\sigma_e^2+n\frac{p-1}{p}\sigma_{\tau b}^2+nm\frac{\sum_j\tau_j^2}{p-1}}{\sigma_e^2+n\frac{p-1}{p}\sigma_{\tau b}^2}$$

需要注意的是，我们以交互项均方而非聚类间均方作为分母。这是因为有效的F检验要求期望分母需包含期望分子中的项(Brown & Melamed，1990)。由于我们已知，若$p=2$，则$n\frac{\sum_j\tau_j^2}{p-1}=\frac{n}{2}(\mu_1-\mu_0)^2$(方程6.6)，且$(p-1)/p=1/2$，所以实际的$F$检验为：

$$F=\frac{\frac{n}{2}m(\mu_1-\mu_0)^2}{\sigma_e^2+\frac{n}{2}\sigma_{\tau b}^2} \qquad [7.6]$$

因此对应的t检验为：

$$t=(\mu_1-\mu_0)\sqrt{\frac{nm}{2\left(\sigma_e^2+\frac{n}{2}\sigma_{\tau b}^2\right)}} \qquad [7.7]$$

表 7.1　干预来源固定、聚类来源随机的双因素方差分析期望均方

来源	期望均方
组间	$E(MS_\tau) = \sigma_e^2 + n\frac{p-1}{p}\sigma_{\tau b}^2 + nm\frac{\sum_j \tau_j^2}{p-1}$
聚类之间	$E(MS_b) = \sigma_e^2 + np\sigma_b^2$
交互	$E(MS_{\tau b}) = \sigma_e^2 + n\frac{p-1}{p}\sigma_{\tau b}^2$
内部	$E(MSW) = \sigma_e^2$

注：MSW 为组内均方。

第 1 节 | 多点随机试验的功效参数

和上一章一样，在不知道测量尺度的详细信息的情况下，我们需要运用无标尺参数来进行功效分析。一种方法是把方程 7.6 中的分子和分母分别都除以 σ_b^2。这样就得到了两个无标尺参数。第一个是效应值除以组内相关系数的平方根：

$$\delta_b=\frac{\mu_1-\mu_0}{\sigma_b}=\frac{\delta}{\sqrt{\rho_{intra}}} \qquad [7.8]$$

另一个是干预效应方差与聚类均值间方差的比值：

$$\upsilon=\frac{\sigma_{\tau b}^2}{\sigma_b^2} \qquad [7.9]$$

把上述两式（方程 7.8 和方程 7.9）与 $\sigma_e^2=1-\rho_{intra}$ 和 $\sigma_b^2=\rho_{intra}$ 这两个假定相结合，我们就能得到一个无标尺的非中心化参数（Hedges & Rhoads，2010）[24]：

$$\lambda=\delta_b\sqrt{\frac{nm}{2\left(\frac{1-\rho_{intra}}{\rho_{intra}}+\frac{n}{2}\upsilon\right)}}=\delta\sqrt{\frac{nm}{\rho_{intra}\left[2\left(\frac{1-\rho_{intra}}{\rho_{intra}}+\frac{n}{2}\upsilon\right)\right]}}$$

因此，非中心化参数为：

$$\lambda = \underbrace{\delta}_{\text{效应值}} \underbrace{\sqrt{\frac{nm}{2}}}_{\text{样本量}} \underbrace{\sqrt{\frac{1}{1+\left(\frac{n}{2}\upsilon-1\right)\rho_{intra}}}}_{\text{设计效应}} \qquad [7.10]$$

多点随机试验中纳入不相关的协变量

方程 7.10 可用于处理协变量问题。假定我们希望在模型中纳入不相关的协变量，譬如前测及其聚类均值。我们把模型进行如下设置：

$$y_{ijk} = \gamma_0 + \gamma_1 T_{ijk} + \gamma_2 (x_{ijk} - \bar{x}_k) + \gamma_3 \bar{x}_k + b_k^* + (Tb)_{jk}^* + e_{ijk}^* \qquad [7.11]$$

我还是用星号（∗）来提醒读者，由于纳入了协变量，随机效应有所下降。方差成分也有所缩减，缩减幅度和协变量与随机效应在实验组内的相关程度有关。譬如，$e_{ik(j)}^*$ 的方差为 $\sigma_e^{2*} = \sigma_e^2(1-R_{unit}^2)$，而 $(Tb)_{jk}^*$ 的方差为 $\sigma_{\tau b}^{2*} = \sigma_{\tau b}^2(1-R_{treat}^2)$。其中 R_{unit}^2 是聚类内方差被协变量所解释的比例，R_{treat}^2 是干预效应的方差被协变量的聚类均值所解释的比例。如下式所示，我们可以把协变量效应整合进非中心化参数中（Hedges & Rhoads, 2010）：

$$\lambda = \delta \sqrt{\frac{nm/2}{1+\left(\frac{n}{2}\upsilon-1\right)\rho_{intra} - \left[R_{unit}^2 + \left(\frac{n}{2}\upsilon R_{treat}^2 - R_{unit}^2\right)\rho_{intra}\right]}} \qquad [7.12]$$

这一检验的自由度为 $m-1-q$，其中 q 是聚类层面协变量的个数。

第2节｜多点随机试验的实例分析

这一节简要分析一些模拟数据。[25]数据可参见附录，是模拟的学生（单位）成绩，取值范围为0—100。成绩很可能受到学校（研究点）内部指定干预的影响。数据包含 $m=5$ 所学校，每所学校包含 $p=2$ 个实验组，每个实验组包含 $n=4$ 名学生。两个混合模型的结果如表7.2所示。

在表7.2第一个不包含协变量的模型中，可见干预条件下的学生成绩平均增加11.34分，但是该效应在统计上不显著。我们可对表中所报告的参数值进行 t 检验。效应值为：

$$\delta=\frac{11.340}{\sqrt{137.670+142.269}}=0.678$$

参数 υ 为：

$$\upsilon=\frac{220.834}{137.670}=1.604$$

而组内相关系数为：

$$\rho_{intra}=\frac{137.670}{137.670+142.269}=0.492$$

由此得到 t 检验（即非中心化参数）为：

$$t=0.678\sqrt{\frac{4\times5}{2}}\sqrt{\frac{1}{1+\left(\frac{4}{2}1.604-1\right)0.492}}=1.484$$

这一结果可以通过系数和其标准误的比值来得到验证，11.340/7.642=1.484。

表 7.2 多点随机模拟数据的模型结果

	混合模型 1	混合模型 2
截距	55.623***	21.435
	(5.886)	(43.852)
干预组对比控制组	11.340	11.340*
	(7.642)	(5.188)
协变量		0.509***
		(0.152)
研究点协变量均值		0.684
		(0.872)
N	40	40
研究点的数量	5	5
σ_b^2	137.670	73.153
$\sigma_{\tau b}^2$	220.834	74.786
σ_e^2	142.269	119.600

注:协变量进行组均值对中。括号中为标准误。* $p<0.05$, ** $p<0.01$, *** $p<0.001$。

表 7.2 中第二个混合模型纳入了不相关协变量,结果显示干预效应没有变化,但在统计上却变得显著了,因为标准误减小了。同时,我们还发现聚类截距和聚类干预效应的方差都减小了。单位截距方差的缩减比例为 $R_{unit}^2=1-\frac{119.6}{142.269}=0.159$,干预效应方差缩减比例为 $R_{treat}^2=1-\frac{74.786}{220.834}=0.661$。因此,现在的 t 检验和非中心化参数为:

$$t = 0.678\sqrt{\frac{4\times 5/2}{1+\left(\frac{4}{2}1.604-1\right)0.492-\left[0.159+\left(\frac{4}{2}1.604\times 0.661-0.159\right)0.492\right]}}$$

$$=2.186$$

这一结果可以通过系数除以其标准误得到验证，11.340/5.188＝2.186。纳入不相关协变量改善了效应估计的精确性，因而干预效应现在变得显著了。

第3节 | 多点随机试验的功效分析

求先验功效

一旦求出了非中心化参数(方程 7.12),功效分析就跟本书中其他 t 检验的分析相同。得到 λ 后,检验的功效就能通过电脑中的 H 函数(即第4章的方程 4.23 和方程 4.24)计算得到。需要注意的是,该检验的自由度为 $m-1-q$,其中 q 是检验中所使用的聚类层面协变量的个数。如果未使用协变量,就可以把 q、R^2_{unit} 和 R^2_{treat} 简单地设为 0。

求必要样本量

和聚类随机设计一样,增加单位数量对于功效改进的效应是不断递减的。然而,需要注意的是,这里的 n 是每个聚类中每个实验组所包含的单位数。因此,如图 7.2 所示,当等高线在 $n=30$ 之后开始变得垂直,这意味着每个聚类包含的总单位数为 60。这表明,与聚类随机试验相比,在多点随机试验中增加聚类的单位数成本更高。尽管如此,我们仍会聚焦于求最低的聚类数量。

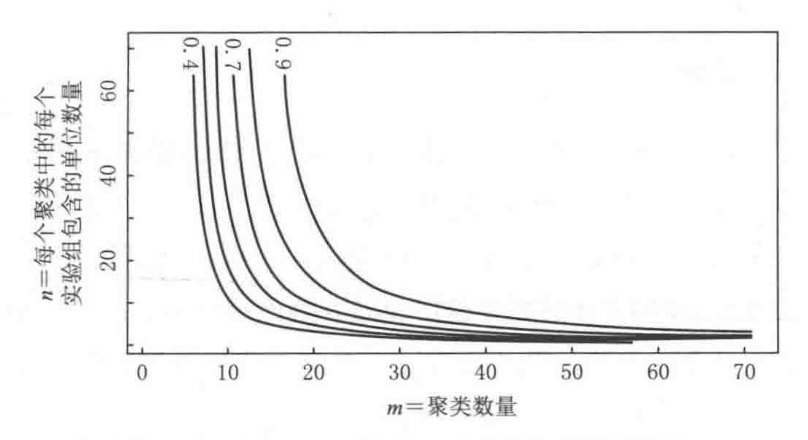


图 7.2　$\delta=0.3$、$\rho_{intra}=0.2$ 且 $\upsilon=0.5$ 的多点随机设计中，m 和 n 不同取值情况下的功效等高曲线图

求多点随机试验中的最低聚类数量的程序与聚类随机试验类似。我们首先用与方程 6.25 类似的方式求正态分布近似值：

$$m_Z \approx \frac{2(z_{critical}-z_{\beta})^2 D}{n\delta^2} \qquad [7.13]$$

其中，D 是设计效应，其取值不同于聚类随机试验中的 D：

$$D=1+\left(\frac{n}{2}\upsilon-1\right)\rho_{intra}-\left[R^2_{unit}+\left(\frac{n}{2}\upsilon R^2_{treat}-R^2_{unit}\right)\rho_{intra}\right] \qquad [7.14]$$

一旦得到了这个近似值，就可以用 t 分布的分位数再次改进我们的估计：

$$m \approx \frac{2[t_{(m_Z-1-q)critical}-t_{(m_Z-1-q)\beta}]^2 D}{n\delta^2} \qquad [7.15]$$

需要注意的是，这里使用的自由度为 $m-1-q$，而非 $2m-2-q$。

实例

譬如,假定我们正在设计一个功效为 0.8 的研究,在 $\alpha=0.05$ 水平上进行双尾检验,期望效应值为 0.25 左右,假定 υ 比值约为 0.25,组内相关系数约为 $\rho_{intra}=0.2$,并使用了一个协变量,该协变量解释单位层面的方差比例约为 $R^2_{unit}=0.5$,解释的干预方差比例约为 $R^2_{treat}=0.35$。若计划每个聚类包含 $n=50$ 个单位,设计效应为:

$$D=1+\left(\frac{50}{2}\times 0.25-1\right)\times 0.2-\left[0.5+\left(\frac{50}{2}\times 0.25\times 0.35-0.5\right)0.2\right]$$
$$=1.213$$

故而聚类数量的正态近似值为:

$$m_Z\approx\frac{2[1.96-(-0.842)]^2\times 1.213}{50\times 0.25^2}=6.095$$

即约需七个聚类。然后把这个数字运用到自由度为 7−1−1=5 的 t 分布近似值中,便得到 $t_{(m_Z-1-q)critical}=t_{(7-1-1)0.975}=2.571$, $t_{(m_Z-1-q)\beta}=t_{(7-1-1)0.2}=-0.92$。因此可求得:

$$m\approx\frac{2[2.571-(-0.92)]^2\times 1.213}{50\times 0.25^2}=9.461$$

即约需 10 个聚类。这一检验的功效实际约为 0.88, 9 个聚类的功效约为 0.83。这说明在小样本情况下,一般来说,稳妥起见,在使用近似方法时最好取较大样本量。

求最低可检测效应值

和以前一样,通过变换可得到最低可检测效应值的

公式：

$$\delta_m = [t_{(m-1-q)critical} - t_{(m-1-q)\beta}]\sqrt{\frac{2D}{nm}} \qquad [7.16]$$

其中，D 是设计效应，与方程 7.14 中的定义相同，而 q 则是协变量的数量，用于求 t 分布的分位数。

实例

假设已有资源仅允许我们包含 $m=7$ 个聚类，则问题便是最小可检测效应会是多少。我们可以使用和以前相同的设计效应（$D=1.213$），然后把 t 分布分位数修正为 $t_{(m_z-1-q)critical}=t_{(5)0.975}=2.571$ 和 $t_{(m_z-1-q)\beta}=t_{(5)0.2}=-0.92$，计划每个学校包含 $n=50$ 个学生，则最低可检测效应为：

$$\delta_m = [2.571-(-0.92)]\sqrt{\frac{2\times1.213}{50\times7}} = 0.291$$

因此效应值接近于我们所希冀的 0.25。

第 4 节 | 小结

本章继续探讨了多层设计,特别是层 1 单位作为随机化单位的两层设计。虽然这类研究的分析通常采用混合回归程序,但运用方差分析框架来进行多层设计的功效计算使读者更易理解。本章使用期望均方,得到了多点随机试验的非中心化参数。比较图 6.2 和图 7.2,我们发现在相同样本量情况下,多点随机试验的功效要大于聚类随机试验。然而,当同一个组织中混合了干预和控制情况时,多点设计更可能遭遇混杂效应的干扰及其他困难(Bloom, 2005)。因此,聚类随机试验虽然需要更多的聚类,但可以通过产生更大的效应值来弥补其较低的功效。



合理的假定

行文至此,功效分析的基本进路已甚为明晰。为了计算目标假设,每种功效分析(求功效、求样本量、求效应值)都会对研究议题的部分内容作出假定。为了求功效,我们就需要假定效应值和样本量。为了计算样本量,我们就需要假定功效水平和效应值。为了计算最低效应值,我们要假定功效和样本量。而且,我们往往还需要假定第一类错误的水平以及检验类型(但一般都使用 $\alpha=0.05$ 水平上的双尾检验,在这一点上研究者无甚争议)。在更为复杂的设计中,我们必须假定更多的参数。对于协变量分析模型,我们需要假定协变量与结果变量之间的相关关系。对于聚类样本,我们需要假定组内相关的情况。由此可见,功效分析皆关乎假定。

从孩提时代起,我们就知道假定会使人看起来蠢笨。由于功效分析完全基于假定,使它成为一种颇为微妙、需要小心处理的方法。为了帮助读者看起来不那么蠢笨,本章为如何设置有效假定提供一些策略。

虽然不可能总是得到一个“正确的”功效分析,但我们的目标应该是使研究基金或合同中所使用的功效分析正当合理。作为一名作者和评审人,我经常发现,虽然没有哪个基金项目的成功仅仅是因为强大的功效分析,但很多项目申请

因为功效分析不行而惨遭淘汰。

本章将探讨如何为研究设计的参数设置合理假定提供具有可行性的策略。我们不想写成一个假定设置过程的辩护型指南，因为这主要取决于研究内容。然而，我们希望这些常识性的实践活动有利于研究者避免常见的陷阱。

第1节 | 功效分析是一种观点

恰如研究者所为,功效分析是基于一组前提假定而得到的一个观点。在争论某个研究是否值得进行投入(如金钱或风险)时,我们的观点是,该研究很可能会成功。当然“成功”有很多含义,但让我们暂时假定成功意味着能对某个正当合理的重要问题给出答案。功效分析对此至关重要,因为它对检测到某种效应的几率进行估计,预估该效应是否真实存在。

功效分析观点围绕设计的两个基本要素来形成假定。首先,样本设计决定了样本类型并对样本所有要素进行量化。就“简单随机样本”而言,重要的量化指标是总样本量。样本设计细节也与分析类型相关。就两组均值差异比较而言,样本设计就涉及样本是否均衡。更为复杂的样本就牵涉更复杂的量化指标。聚类样本包含聚类的样本量和每个聚类内部的样本量。如果对聚类进行分组,那么就需要确定每个组所包含的聚类数量,等等。大多数的样本设计都由研究小组掌控(或受其限制)。[26]

第二类功效分析观点围绕设计参数,譬如期望效应值和相关性。前述的抽样设计指定了为评估功效所必需的设计参数。比如,在简单随机样本中,组内相关系数就不是必需

的,但若分析是为了比较两组或更多组别,那么效应值就是必需的。

不幸的是,很多研究者易滑入一个陷阱中,即坚持某个抽样设计(尤其是样本量),然后求设计参数以支持样本设计。这类似于选择那些只支持某个特定观点的证据。这往往导致不合理的期望,即效应值或相关性较高,但实际上它们可能都比较小。

"多大算大?"

诚然,"小"和"大"都是相对的,某个数值是大是小在不同学科、不同干预和不同情境之间存在很大差异。以往的情况是,很多研究者都依靠科恩(Cohen, 1992)提出的所谓"衬衣尺码"(shirt sizes)方法。[27] 仰赖于这样一个由 20 世纪的一群自然科学家事先设定的量化指标是危险的。科恩自己也表示,盲从这些规则对于更广大的科学领域并无裨益。之所以危险,乃是由于使用了一个武断的数值,回避了什么构成了与总体变异相关且有意义的差异这一重要问题。虽然有些软件仍支持设定这类衬衣尺码值,但我们应忽略它们,或仅作为承认无知的无奈之举。

越来越多的研究开始专门估计真实发生的效应值,并以之为标准来评估某个干预导致的变化。一个实例是希尔等人估计与每接受一年教育回报相关的效应值(Hill et al., 2008)。当然,这取决于多个因素,包括年龄和结果变量。在阅读表现上,从幼儿园到一年级(一年的教育年限)增加约 1.5 个标准差,在数学表现上增加 1.14 个标准差。之后,

从 11 年级到 12 年级,阅读表现增加 0.06 个标准差,数学表现增加 0.01 个标准差(Hill et al., 2008)。这一研究对教育干预非常有帮助,因为它在教育研究中提供了一种直观印象,给出了可能效应值的上限。[28]

假设某研究者在一项基金申请中提出要评估某个旨在增进学业表现的课程。该研究者认为期望效应值为 0.1。按多个标准来看,这个效应较"小",看起来也较为合理。若研究者在幼儿园学童的阅读表现上使用这一干预,0.1 个标准差的增量约代表 7%的一年教育回报*,我们就可以认为这是合理的。但倘若这一干预是针对 11 年级学生的,0.1 这一"小"效应则是一年教育回报量的 10 倍,那么这一假定显然就不合理了。

合理的假定是合理功效分析的基石。获得合理期望的最好办法是利用已有研究的数据或试调查数据,因为这将很接近当前研究的情况。在多数情形中,无法获得这一类数据,那就只能依靠文献。本章的剩余部分将提供一些指南,探讨如何利用已有研究来形成功效分析的假定。

---

* 7%=0.1/1.5。作者在这里的意思是干预效应的大小是否合理取决于不同年龄段或年级段。——译者注

第2节 | 利用文献形成合理假定的策略

这一部分介绍利用文献形成假定的一些策略。这些策略更像是艺术而非科学，涉及很多研究者的主观判断。搜寻过程中的一项关键技能是从研究中收集重要统计量用于计算我们感兴趣的设计参数。这里所述的诸多技能由元分析研究者发展而来(Borenstein, Hedges, Higgins, & Rothstein, 2009)。事实上，对于收集效应值或相关系数而言，元分析是一个不错的起点。

均值和相关性中的标准化差异

对于旨在探索变量间关系的任何研究而言，其功效分析过程中最重要的假定是估计两个变量之间的相关性，无论是组均值差异还是线性相关。即使在估计最低可检测效应时，对于比较分析而言，非常重要的一点是了解什么是可能的。在文献中搜寻与结果变量、干预条件或与两者都相关的证据总是“磨刀不误砍柴工”。这会避免使用“衬衣尺码”方法所导致的缺陷，因为“衬衣尺码”法仅在某个领域、针对某个特定总体考察效应值大小。

遗憾的是，除非研究目的是复制某项研究，否则研究者不太可能找到一个在相似的结果变量上使用相似干预的研究。[29]因此，在看其他实验研究、考察文献中的结果时，我们需要平衡两个问题：

（1）自变量存在多大相似性？

（2）因变量存在多大相似性？

对设计参数的指引主要来自对不同研究的结合，这些研究有的存在相似的干预（这可以告诉我们某个干预的有效性如何），有的存在相似的结果变量（告诉我们某个干预对结果变量的效应幅度如何）。这两个视角对效应值的确定都有所裨益。遗憾的是，哪个视角更重要仍然是一个更为艺术性而非科学的问题。

元分析是一个不错的出发点，因为元分析经常报告效应值。当找不到元分析时，文献检索能发现干预变量或结果变量相似的研究。一旦找到了相关研究，我们就能用效应值的公式提取所需的设计参数。若某篇文章提供了每个组的样本量、均值和标准差，我们就能计算标准化均值差异。若文章提供了相关性系数，而非标准化均值差异，我会在后面提供一个转化公式。在实际操作中，研究者应该收集尽可能多的研究来组装他们的假定。下面的例子中我仅用了两个研究，这是出于简洁性的考量，在实际研究中应当广撒网。

实例

试看下面这个例子。考察警察的穿戴式摄像头对警务活动各方面影响的文献在不断增长（如 White, 2004）。假定

某个研究组计划让警官使用穿戴式摄像头，考察这是否会增加市民对警察合法性的认可。假定之前未进行过此类干预性研究[30]，那么研究者就只好考察两方面的文献：使用干预（穿戴式摄像头）的研究，或考察结果变量（警察合法性）的研究。

有一项研究在佛罗里达州随机指定使用穿戴式设备，以检验其对民众对警官的严重投诉的效应（我们可以认为这和合法性感受相关；参见 Jennings，Lynch，& Fridell，2015）。这一研究中，控制组包含 $n_0=43$ 名警官，干预组包含 $n_1=46$ 名警官。控制组的投诉平均数为 0.19，标准差为 0.39。[31]干预组的投诉平均数为 0.09，标准差为 0.28。回忆一下，效应值是一个无标尺参数，是组均值的标准化差异，$\delta=\frac{\mu_1-\mu_0}{\sigma}$。我们可以计算合并标准差，这是 σ 的估计值，即方程 4.6 的平方根，

$$\hat{\sigma}=\sqrt{\frac{(43-1)0.39^2+(46-1)0.28^2}{43+46-2}}=0.34$$

由于差异约为 0.1 个投诉，所以标准化均值的效应值约为$\delta=0.1/0.34=0.29$。

另一个研究评估了澳大利亚的一项干预，考察若给警官提供一个旨在增加程序正义感受的脚本，是否会对市民遇到警察后的感受有所影响（Mazerolle，Antrobus，Bennett，& Tyler，2013）。随机选择警官让他们按照旨在增加合法性感受的脚本行事，或按照以往方式行事，然后让他们在事先设置的路障处遇到市民。虽然这篇文章使用了复杂的分析，但还是提供了重要的信息可用于确定科恩的 d 效应值。干预

组被访者的合法性量表(取值为1—5)均值为4.21,控制组被访者的均值为3.96。[32]两组的标准差皆为0.72左右。[33]因此,效应值约为(4.21−3.96)/0.72=0.35。

在这个例子中,我们发现干预对合法性感受的效应值约为0.35,穿戴式摄像头对投诉的效应值约为0.29。没有一个明确答案告诉我们应该用哪一个数值作为我们的效应值。然而,既然这两个数都接近于0.3,那么就促使我们思考,在我们所要开展的实验中,一个合理的效应值或许是0.3。[34]当然,更多的文献检索或元分析对于确定合理的效应值有很大帮助。

把相关性转化为均值差异

观察性研究或实验研究的结果通常会提供相关性系数以概括某个干预的效应。在其他情况中,某个干预或许会利用已被证明与结果变量相关的某个变量。在这些情况中,通常需要把相关系数转化为科恩的 d 效应值。元分析文献在此颇有裨益,因为元分析经常整合来自不同研究的效应,这些效应往往使用不同的度量(Borenstein et al., 2009)。[35]

为了把相关系数(ρ)转化为标准化均值差异(如科恩的 d 或赫奇斯的 g),我们可以使用下述公式(Borenstein et al., 2009):

$$\delta=\frac{2\rho}{\sqrt{1-\rho^{2}}} \qquad [8.1]$$

例如,若某篇文章指出干预变量和结果变量间的相关系数为0.18,恰如澳大利亚那个研究中所言(Mazerolle et al.,

2013)，我们就可以把它转化为标准化均值差异：

$$\frac{2\times0.18}{\sqrt{1-0.18^2}}=0.4$$

因此，0.18的相关性等同于均值差异为0.4个标准差。如读者所见，这一结果不同于我们从概要统计中计算得到的0.35。之所以不同，是因为这些公式简化了复杂关系，故而在求可能的效应值过程中应仅作为参考。效应值的相关系数的反向转化公式为(Borenstein et al., 2009)：

$$\rho=\frac{\delta}{\sqrt{\delta^2+a}}$$

其中，a是每组样本量的一个函数，

$$a=\frac{(n_0+n_1)^2}{n_0n_1}$$

如果该项研究是均衡设计的，那么$a=4$。譬如，我们可以利用文章所提供的样本量($n_0=1\ 649$, $n_1=1\ 097$)和效应值0.35，把警察合法性研究中的标准化均值差异转化为相关系数：

$$\frac{0.35}{\sqrt{0.35^2+(1\ 649+1\ 097)^2/(1\ 649\times1\ 097)}}=0.17$$

这非常接近该文所报告的相关系数0.18。

复杂样本的设计参数

现在越来越多的研究涉及复杂样本，因而打破了简单分析所要求的假定。最常见的情况是聚类样本[36]，我们感兴

趣的效应来自聚类层面变量(譬如聚类随机试验)或聚类内部变量(譬如多点随机试验)。

在第6章,我们知道效应之抽样方差的差异可以被概括为“设计效应”。例如,某聚类随机试验的设计效应如方程6.21所示,$D=1+(n-1)\rho_{intra}$。其中,n 是每个聚类所包含的单位数量;ρ_{intra} 是组内相关系数,以测量某聚类内部单位之间的相关程度。譬如,若 $\rho_{intra}=0.1$,且每个聚类包含30个单位,则组均值差异的方差就会增加 $1+(30-1)0.1=3.9$。这意味着标准误(方差的平方根)也会增加,从而导致第一类错误增加,即更不显著。

第7章也详述了多点随机试验(即效应来自聚类内部的某个变量)的设计效应。这里,除了组内相关之外,设计效应还涉及不同聚类间干预效应的方差。设计效应如方程7.10所示,

$$D=1+\left(\frac{n}{2}\upsilon-1\right)\rho_{intra}$$

其中,υ 是聚类间干预效应方差与聚类均值方差的比值(详见第7章)。

计划运用复杂样本的研究一般从简单随机样本的解决方法开始,然后再使用合适的设计效应。例如,若运用简单随机样本的期望 t 检验统计量(即非中心化参数)为2.5,但计划中复杂样本的设计效应预计为3.9,那么我们就必须把非中心化参数修正为 $2.5/\sqrt{3.9}=1.27$。这使功效大大缩减,因为非中心化参数减小了很多。因此,复杂样本中的设计效应预判非常重要。

虽然设计效应受聚类内部样本量(n)这一研究者可掌控

因素的影响，但也受到组内相关(ρ_{intra})和干预异质性参数(比如 υ)的影响。而后两个参数往往是未知的。

健康和教育研究领域的学者已经做了很多工作，对可用于计划研究的总体组内相关估计值进行了编目整理。具体实例包括运用社区健康调查的研究(Gullifor, Ukoumunne, & Chinn, 1999)、学校药物滥用研究(可参见 Murray et al., 1994)、教育实验(Schochet, 2008)、全国学业成绩调查(Hedges & Hedberg, 2007)，以及各州的教育系统成绩记录(Hedberg & Hedges, 2014; Hedges & Hedberg, 2013; Westine, Spybrook, & Taylor, 2013)。多数这些研究包含了与通常的协变量组相关的 R^2，因而研究者就能把前测和人口控制变量用于预估有效性。

在健康研究中，默里及其同事(Murray et al., 1994)制作了一组青少年抽烟嵌套于学校的组内相关系数，可用于未来抽烟研究的聚类随机试验。他们著作中的表 6 提供了包含不同年级的大量研究中每周抽烟数量的组内相关系数(Murray et al., 1994)。从 11 年级到 12 年级，组内相关系数是 0.076，意味着计划包含 $n=30$ 的某个研究所期望的设计效应为 $1+(30-1)0.076=3.204$。这一设计效应说明，该研究期望效应的抽样方差是简单随机样本的三倍多。

教育是另一个广泛使用此类设计参数的研究领域。过去十年间，赫奇斯(Hedges, 2007, 2013)和赫德伯格(Hedberg, 2014)，韦斯廷、斯皮布鲁克、泰勒(Westine, Spybrook, & Taylor, 2013)以及其他学者发表了一组文章，提供了涉及不同年级、不同学科的学业成绩的组内相关系数。例如，全国幼儿园数学成绩的组内相关系数是 0.243(参见 Hedges &

Hedberg, 2007 的表 2)。这意味着计划包含 $n=30$ 的某个研究所期望的设计效应为 $1+(30-1)0.243=8.047$。这一设计效应说明,该研究期望效应的抽样方差是简单随机样本的八倍多。

我们对多点间效应的异质性所知甚少,故而提供 v 之类参数值的文章所见寥寥。一般来说,要估计这一参数,最好手头有一些数据。

第3节 | 小结

不得不承认，本章就功效分析所提出的问题要多于所提供的答案。换言之，功效分析几乎完全建立在效应值和聚类内部相关这一类假定之上。研究者永远无法很肯定地预估未来数据中的这些参数，不过本章所要表达的是，去猜测这些假定或用常规做法来代替猜测皆非明智之举。更好的办法是从文献检索中寻找预估的线索。若存在与研究主题相关的元分析，这将会是一个有利资源。更常见的情形是不存在这类研究，那么就必须把不同来源的证据拼凑起来。我们很难得到一个确定不移的答案，但它至少能提供一些信息。

第9章

功效的报告

本章讨论如何报告功效,目的是让研究者了解一项细致的功效分析必须报告哪些内容。基金申请和合同计划书中有关功效分析的版面通常非常有限,因此必须谨慎对待评审人所希望看到的每一项内容,评审人就根据这几段内容对研究计划书作出判断。

第1节 | 包含的内容

第8章指出功效分析是种观点，应当基于合理的假定。诚然，申请书中有关功效分析的部分应包括所有的假定。然而，这还不够，因为功效分析还必须说明检测效应的功效是如何计算的，而且是以一种与研究设计相契合的方式。譬如，某设计需要聚类样本，那么细致讨论简单随机样本的功效分析就不太明智。接下来，我们就功效分析写作中需要包含哪些内容提供一些建议。

功效分析部分应包含的要素

功效分析写作应根据功效分析的类型不同而有所不同，分析类型包括求功效、求样本量或求最低可检测效应值。一般而言，功效分析需包含如下要素：

(1) 说明所计划的样本设计（如简单随机样本、聚类样本等）和数据结构；务必在设计中包含影响最终样本量的删截或缺失等因素。

(2) 说明用于效应估计的统计模型，以及是否使用协变量。

(3) 说明功效分析中所使用的设计参数假定，包括支持

这些假定的引用文献和数据;务必包括假定的第一类错误水平(α)以及单尾还是双尾检验(若适用)。

(4) 说明用于功效分析的公式(包括表达式和页码)和软件(包括程序或命令)的出处,以及它们如何与统计模型相关联。若功效分析用到了某个一般性的统计软件包,则有必要提供一个包含代码的附录。

(5) 下述分析结果可在这一部分报告:若分析旨在估计特定样本量和效应值情况下的功效,则应报告所得的功效水平。若分析旨在估计特定功效水平和效应值情况下的样本量,则应当报告最低样本量,以及预算和招募相关事宜。若分析旨在估计特定样本量/样本设计与功效水平情况下的最低可检测效应值,那么就应报告效应值与以往文献的比较结果。

(6) 报告偏离设计参数假定或抽样设计假定的敏感性分析结果,若版面空间允许,最好画图。

(7) 说明用于形成假定的研究总体能有效拟合待研究总体(可选项,但正变得日益重要,因为外在效度日益受到关注)。

第 2 节｜实例

在这一节，我会用一些实例来演示如何包含上面所提及的所有要素。第一个例子针对第 4 章讨论的简单随机样本。第二个例子针对第 6 章所讨论的较为复杂的样本。这些例子应仅仅被当作例子，因为不同的资助机构、委托人、期刊、专业协会或伦理审查委员会可能都会有专门的功效分析写作指南。

我们的例子还是回到第 8 章所论及的穿戴式摄像头研究。倘若我们正在计划一个考察穿戴式摄像头如何影响警察合法性感受的研究。在申请书的其他部分，我们详述了数据收集过程和数据分析。需要说明的是，这两个例子都是简化版本。这样的研究在实际操作中更复杂，譬如需要考虑警察小组和巡逻区域等。

简单随机样本

在这两个例子中，假定调查问卷是发放给遇到这 10 个佩戴摄像头或 10 个未佩戴摄像头的警官的市民。对于每位警官，研究小组已确定研究期间会回收约 25 份邮寄问卷。问卷是匿名寄给研究小组的，然后输入数据库中。我们使用

简单 t 检验来比较干预组和控制组的警察合法性分数。为简便起见,我们假定样本是均衡的。

给定效应值和样本量情况下的功效

下面的段落是概述所得功效的可能方式:

> 我们计划从城市警局中抽取 20 名警官,然后随机选取 $n=10$ 名警官使用穿戴式摄像头(干预组),$n=10$ 名警官不使用穿戴式摄像头(控制组)。在研究期间,每位警官出外勤会遇到约 100 位接受问卷调查的市民。我们预计问卷回收率约为 25%,从而得到的总样本量为 $2\times10\times25=500$。均值比较的分析方法是对问卷的合法性量表分数进行独立样本 t 检验。已有研究显示,在佛罗里达州,穿戴式摄像头会影响市民投诉约 0.29 个标准差(计算过程见附录 A;参见 Jennings et al., 2015);澳大利亚的一项研究发现,有关警察行为的干预会对市民的警察合法性反馈产生 0.35 个标准差的影响(计算过程见附录 A;参见 Mazerolle et al., 2013)。因此,我们假定科恩的 d(Cohen, 1988)效应值约为 $\delta=0.3$。运用 G*Power 软件包中的"两个独立均值间差异"程序(Faul, Erdfelder, Lang, & Buchner, 2007),针对我们的样本和效应值,计算得到的功效为 0.92。这一结果对效应值敏感,譬如,检测出 0.25 个标准差的功效是 0.80。这一结果对问卷回收情况更为敏感,譬如,若每位警官仅回收 10 份问卷,那么检测出 0.3 个标准差的功效约为 0.56。

给定功效和样本量情况下的最低可检测效应值

下面的段落是概述最低可检测效应值的可能方式：

我们计划从城市警局中抽取20名警官，然后随机选取 $n=10$ 名警官使用穿戴式摄像头（干预组），$n=10$ 名警官不使用穿戴式摄像头（控制组）。在研究期间，每位警官出外勤会遇到约100位接受问卷调查的市民。我们预计问卷回收率约为25%，从而得到的总样本量为 $2\times10\times25=500$。均值比较的分析方法是对问卷的合法性量表分数进行独立样本 t 检验。由于给定的所需功效为0.8，我们计算的最低可检测效应值以科恩的 d（Cohen, 1988）来度量。运用G*Power软件包中的“两个独立均值间差异”程序（Faul et al., 2007），针对我们的样本量和所需功效（0.8），计算得到的最低可检测效应值为0.25。已有研究显示，在佛罗里达州，穿戴式摄像头会影响市民投诉约0.29个标准差（计算过程见附录A；参见Jennings et al., 2015）；澳大利亚的一项研究发现，有关警察行为的干预会对市民的警察合法性反馈产生0.35个标准差的影响（计算过程见附录A；参见Mazerolle et al., 2013）。因此，我们有足够的信心认为我们的样本足以检测到与文献相关的效应。这一结果对样本量敏感，譬如，若每个警官仅回收10份问卷，则最低可检测效应就会更大，约为0.40。

给定功效和效应值情况下的最小样本量

下面的段落是概述最小样本量的可能方式（需要注意的

是我们想要求的是，这 20 位中的每一位警官出外勤所遇到的被问卷调查的市民数量）：

> 我们计划从城市警局中抽取 20 名警官，然后随机选取 $n=10$ 名警官使用穿戴式摄像头（干预组），$n=10$ 名警官不使用穿戴式摄像头（控制组）。均值比较的分析方法是对问卷的合法性量表分数进行独立样本 t 检验。由于给定的所需功效为 0.8，我们计算科恩的 d（Cohen，1988）效应值，预计为 0.3。已有研究显示，在佛罗里达州，穿戴式摄像头会影响市民投诉约 0.29 个标准差（计算过程见附录 A；参见 Jennings et al.，2015）；澳大利亚的一项研究发现，有关警察行为的干预会对市民的警察合法性反馈产生 0.35 个标准差的影响（计算过程见附录 A；参见 Mazerolle et al.，2013）。运用 G*Power 软件包中的“两个独立均值间差异”程序（Faul et al.，2007），我们计算得到每个实验组的最小样本量为 176。因此，我们需要平均每位警官有 18 份问卷调查。这一结果对效应值敏感，譬如，若效应值为 0.2，则每位警官需要收集 40 份问卷调查。

给定效应值和样本设计情况下的聚类随机试验的功效报告

为简便起见，我在这一部分仅展现给定样本设计和效应值情况下求功效的功效分析。这个例子假定调查问卷是发放给遇到这 125 个佩戴（干预组）或 125 个不佩戴（控制组）穿

戴式摄像头的警官之市民。对于每位警官，研究小组已确定研究期间会回收约25份邮寄问卷。问卷是匿名寄给研究小组的，然后输入数据库中。我们将使用多层混合模型来比较嵌套于警官的调查问卷结果，这些警官或属于干预组，或属于控制组。为了分析便利，我们在此假定样本是均衡的。

假定我们收集了试调查数据来估计针对每位警官的市民调查问卷的内部相关程度，结果显示组内相关系数为 $\rho_{intra}=0.7$。试调查分析还显示协变量解释了市民层面1/4的方差（$R^2_{unit}=0.25$）和警官层面1/10的方差（$R^2_{cluster}=0.1$）。协变量包括市民调查层面的接触环境（如交通临检点）和警官层面（如在交通临检点的接触比例）的 $q=15$ 个虚拟变量。基于这些参数，功效分析发现，若我们给干预组和控制组各配置125位警官[37]，那么功效会达到0.8。

下面的段落是概述所得功效的可能方式：

> 我们计划从城市警局中抽取250名警官，然后随机选取 $n=125$ 位警官使用穿戴式摄像头（干预组），$n=125$ 名警官不使用穿戴式摄像头（控制组）。在研究期间，每位警官出外勤遇到约100位接受问卷调查的市民。我们预计问卷回收率约为25%，从而得到的总样本量为 $2\times125\times25=6\ 250$。我们将选取适用于调查问卷嵌套于警官的聚类随机试验分析方法。具体而言，分析方法将采用混合回归模型来估计警官组别（干预对比控制）对调查问卷中合法性量表的效应。已有研究显示，在佛罗里达州，穿戴式摄像头会影响市民投诉约0.29个标准差（计算过程见附录A；参见Jennings et al.，2015）；

澳大利亚的一项研究发现,有关警察行为的干预会对市民的警察合法性反馈产生0.35个标准差的影响(计算过程见附录A;参见Mazerolle et al., 2013)。因此,我们假定科恩的d(Cohen, 1988)效应值约为$\delta=0.3$。

试调查数据显示嵌套于警官的调查问卷的组内相关系数为0.7,但包含30个与市民接触环境相关的协变量(其中15个属于市民调查问卷层面,15个属于警官层面)解释了调查问卷层面25%的方差和警官层面10%的方差。根据多层设计的公式(参见Hedges & Rhoads, 2010中第21页的表达式9),利用这些参数和Stata中的RDPOWER软件(Hedberg, 2012),我们得出,在$\alpha=0.05$水平的双尾检验中,这一设计的功效为0.84。

这一结果对效应值敏感,譬如,检测到0.25个标准差的功效约为0.69。该结果对调查问卷回收数量不是那么敏感,譬如,若每位警官仅回收10份调查问卷,检测到0.3个标准差的功效约为0.83。

第3节 | 小结

本章建议了功效分析报告所需的要素。我们回到了第8章中有关穿戴式摄像头的研究实例,写了一段假想性的功效分析段落。需要再次说明的是,列举的要素和样板语言仅提供指南,因为不同基金或评审申请会有不同的要求。最佳选择就是完全按照申请说明来写功效分析报告。



第10章

结论、拓展阅读和回归

作为功效分析的概论,本书的核心是概述样本设计(如样本量)与设计参数(如效应值)之间的互动。功效分析整合了这些要素,试图估计出检测到某个效应的几率(假如这一效应存在的话)。功效分析还能假定检测到效应的几率,然后去估计最小样本量或最低可检测效应。本书的任务便是以两组间均值差异这一常见分析为例,来探讨功效分析的机制。

本章试图把本书中的所有内容联系起来。我阐述了聚焦于两个组别和随机化的理由,然后提供一些有助于拓展至其他功效分析主题的阅读资料。最终,我们的最后一个练习是讨论观测数据回归设计的功效。

第 1 节 | 比较两个组别的个案研究

为了对功效分析进行简介，本书考察了不同复杂水平的两组比较检验，从简单随机样本到更为复杂的聚类设计。这两类设计都在回归分析框架下考虑到了协变量的使用问题。对于没有协变量的简单随机样本，我们还考虑了均衡问题，即两个组别是否包含相同的观测数。

这部分的目的不是对每种需要进行功效估计的分析进行包罗万象式的回顾。此外，即使对于简单的两组比较，本书也未讨论所有相关问题。如前文脚注所言，本书并未考虑运用分层样本、抽样权重，或其他检验干预效应的技术，如匹配或倾向值，以及其他方法，等等。

然而，本书的主旨是向读者介绍任何功效分析都不可或缺的关键要素：理解检验统计量及其抽样方差的需求、控制条件的效应，以及偏离简单随机样本的设计效应概念。倘若读者希望使用加权的分层聚类样本，诸如沃尔特（Wolter，2007）等人的书说明了如何理解不同设计的抽样方差对设计效应的影响，而且这些设计效应也可以运用到本书所论及的简单随机样本。其他的重要因素包括删截和未应答，因为这会影响进入分析的数据总量（除非研究组使用插补，当然这会产生一个新的设计效应）。

第2节 | 拓展阅读

我尝试在本书的多个地方都引介相关的拓展性读物。在已有综述的基础上,我再提供下述建议。很多旨在理解随机化设计和观测性设计的功效、样本量和设计效应的研究目前正在进行中。多点之间干预效应的方差是一个仍在不断发展的研究领域(Rhoads, 2017),而我们对于把聚类设计应用到回归模型中已经有了更深入的理解(Lohr, 2014)。抽样和样本量领域的一本经典著作是基什(Kish, 1965)的作品,而更晚近的著作则是瑞安(Ryan, 2013)的作品。这一领域的主要教科书是科恩(Cohen, 1988)这本影响深远的著作。虽然本书对效应值预估的"衬衫尺码"方法持批判态度,但我们也应注意到科恩自己也对该方法的广泛使用持批判态度,并认为自己对此负有责任(如 Cohen, 1994)。

本书批评了复杂设计中回归分析的诸多问题,从而转向了方差分析框架。但回归分析框架有助于更好地理解功效的参数。我鼓励读者去读不使用线性代数的老旧社会统计学教科书,以理解如何从双变量相关到多元回归模型的建构[个人最喜欢布莱洛克(Blalock, 1972)的著作]。

这些资源也可以把其他设计纳入视野。譬如,断点回归(Bloom, 2012)是颇受欢迎的随机化替代方案,因为它使用

了一个连续的标度变量来设置干预，然后拟合一个包含标度变量和设定变量的多元模型来估计干预效应。这一干预效应本质上是一个匹配估计值，比较了断点之前和断点之后的差异。然而，这类设计的功效异常复杂，因为标度变量与干预设定变量相关（Schochet，2009）。了解典型回归中的双变量关系有助于读者理解断点回归之类设计中的功效分析。

非线性模型

本书主要关注线性结果。但并非社会科学中的所有因变量都是线性的，一些是二分类变量（如选择做某事或不做某事），其他一些是计数变量（如孩子的数量），等等。由于不服从正态分布，非线性因变量的模型必须使用一般线性模型来拟合（McCullagh & Nelder，1989）。但不同于正态分布变量，非线性变量的方差是该变量均值的函数，而正态分布变量的方差与均值是不相关的。

均值与方差间的关联给功效分析带来了问题，因为变量效应会影响均值，从而影响方差。有很多著作旨在解决功效和样本量分析中的这个问题，既包括单层模型（如 Hsieh，Bloch，& Larsen，1998），也包括多层模型（Moerbeek，Van Breukelen，& Berger，2001）。

复杂模型

实际研究中存在的模型和问题当然远比两个组别间的简单比较复杂。目前已经解决了结构方程模型（参见 Mac-

Callum, Browne, & Sugawara, 1996)、生存分析模型(如Lachin & Foulkes, 1986)、中介模型(Fritz & MacKinnon, 2007)以及很多其他模型的功效和样本量估计问题。搜寻这类信息时的一个策略是同时设置"功效"和"样本量",因为不同学科会从不同角度来使用这些软件。另一个建议是跳出你所在的特定研究领域,更有创造性地去思考任何研究问题。一项针对惯犯研究的功效分析可能会得益于医学文献中有关药物检验的功效分析,两者有可能都是纵贯离散结果变量。

成本问题

数据收集需要花钱。我发现自己的研究若涉及一手数据收集,则大部分项目经费都会用于数据收集,仅有一小部分用于数据分析。这就使功效分析变得非常重要,因为涉及了几乎所有的资源。在随机化试验中,无论是单层还是多层设计,了解数据收集的成本结构至关重要。很多时候,一个干预单位的成本远高于控制单位,因为你对干预单位投入更多,比如需要提供某个课程。对于单层模型,只要了解干预和控制单位的成本,以及不使用均衡设计的影响,就有助于决定把多少单位分配到干预或控制条件中。

对于多层模型,劳登布什及其同事已经在教育领域做了很有影响力的设计(Raudenbush, 1997; Raudenbush & Liu, 2000)。需要考虑的因素不仅包括干预和控制单位的成本,还包括招募聚类本身的成本,这一成本因干预和控制条件而有所不同,譬如为了保持控制组学校,可能要花费更多成本。

类似 Optimal Design(Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2006)这样的程序使用与功效计算相同的设计参数,来帮助计算干预组和控制组所应包含的最佳单位数与聚类数。

第 3 节 | 观测数据回归分析

在功效分析的最后冲刺阶段,我简要介绍一下观测数据的功效分析。本书选择随机化试验是因为它们(被期望)可以消除所有自变量和控制变量之间的线性相关,使旨在检测效应的功效计算变得更为简单。如第 5 章所见,消除这一类相关使计算和假定都变得更为容易。接下来我们会看到,观测研究的功效分析会更为困难,因为这相当于要求研究者在数据收集之前就要知道数据的相关性结构。

但这并不是说阻止研究者在观测研究中进行功效分析。举个简单的例子,我们就考虑线性回归。假定模型设置正确,来自简单随机样本的观察数据可以用一般最小二乘法,用一元回归模型或多元回归模型进行拟合。实际上,第 4 章的检验就是预测变量为二分变量的简单一元回归模型,第 5 章的检验是包含一个二分预测变量和一个连续控制变量的多元回归。这两个功效分析都聚焦于二分预测变量的斜率。

但其实功效计算适用于任何斜率、整个模型或特定的斜率组合。这些功效计算可以直接使用 F 分布,所以或许你需要再回顾一下第 2 章。[38] 为了讨论的简洁性,我们仅考虑无标尺参数,如相关系数或多元相关系数。我们鼓励读者去阅读更进阶的著作(如 Liu, 2013)。

一元回归

一元回归模型中，回归斜率的检验等同于测量两个变量相关性的相关系数检验（如 y 和 x 的相关系数 ρ）。因此，我们可以对相关系数进行 F 检验：

$$F=\frac{\rho^2(N-2)}{1-\rho^2} \qquad [10.1]$$

这一 F 检验的分母自由度为 1，分子自由度为 $N-2$。譬如，考虑下面的取值情况：

y	x
4	2
6	9
7	8
5	6

我们计算得到 $N=4$，相关系数约为 $\rho_{yx}=0.876$。则该相关对应的 F 检验约为

$$F=\frac{0.876^2(4-2)}{1-0.876^2}=6.598$$

这一检验（分母自由度为 1，分子自由度为 $N-2=2$）对应的 p 值约为 1.24。我们还可以计算另一个效应值作为这一 F 分布的非中心化参数。该效应值为 f^2（Cohen，1988），其表达式为：

$$f^2=\frac{\rho^2}{1-\rho^2} \qquad [10.2]$$

本例中的非中心化参数是效应值与样本量的乘积：

$$\lambda = f^2 N \qquad [10.3]$$

我们可以用这一公式在非中心化 F 分布中求非中心化参数(λ)。非中心化 F 分布非常类似非中心化 t 分布，唯一差别在于前者是单尾的。犹如我们在 t 检验中所使用的 H 函数(方程 4.23)，我们也可以使用一个类似的函数来求 F 分布的功效：

$$\beta = G[F_{(df_1, df_2)1-\alpha}, df_1, df_2, \lambda] \qquad [10.4]$$

其中，$G[a, b, c, d]$是分子自由度为 b、分母自由度为 c 的非中心化 F 分布在点 a 的累积分布函数，其非中心化参数为 d。点 a 是分子自由度为 df_1、分母自由度为 df_2 的 F 分布的临界值。在上述例子中，自由度 $df_1=1$、$df_2=2$ 的 F 分布在 $\alpha=0.05$ 水平上的临界值约为 18.513。参见附录可知，这等于 $t_{(2)0.975}=4.303$ 的平方。

根据效应值 $f^2=\frac{\rho^2}{1-\rho^2}=\frac{0.876^2}{1-0.876^2}=3.299$ 和样本量 $N=4$，可得到非中心化参数 $\lambda=3.299\times N=13.196$，再利用计算机，得到该检验的第二类错误约为 0.499，则功效约为 0.501。这一程序可用于检验斜率、相关系数和整个模型的功效，而且结果都一样。

全回归模型

全回归模型的拟合度可由多元相关系数 R^2 概括，即报告了结果变量的方差由模型设置的协变量所解释的比例。

基于一组包含 q 个变量的预测变量，多元相关系数 R^2 的统计检验公式形式上与方程 10.1 非常相似：

$$F=\frac{R^2q}{(1-R^2)(N-q-1)} \qquad [10.5]$$

该检验的分子自由度为 q，分母自由度为 $N-q-1$。因此，如果研究者大概了解他们数据中的 R^2，那么就能估计其设计的功效。和之前一样，我们根据自由度得到 F 检验的临界值。然后，效应值公式和单个预测变量的情况完全一致：

$$f^2=\frac{R^2}{1-R^2} \qquad [10.6]$$

譬如，假定我们相信某个包含 3 个预测变量、15 个观测值的模型 R^2 会是 0.25。那么 F 检验在 $\alpha=0.05$ 水平上的临界值为 $F_{3,11}=3.587$。其效应值为 $f^2=\frac{0.25}{1-0.25}=0.333$，从而非中心化参数为 $\lambda=0.333\times15=4.995$，因此相关的第二类错误为 0.679，功效为 0.321。

一组预测变量

假定某研究组有个很好的想法，认为一组控制变量 c 能解释一部分因变量变异，记为 $R^2_{control}$。但研究组想要检验增加了变量组 k 之后（和控制变量 c 一起）是否能解释更大比例的因变量变异，记为 R^2_{total}。对这一新变量组（k）的检验还是用 F 检验：

$$F=\frac{(R^2_{total}-R^2_{control})/k}{(1-R^2_{total})/(N-k-c-1)}$$

这一检验的分子自由度为 k，分母自由度为 $N-k-c-1$。因此，若研究者大致了解数据中的 R^2_{total} 和 $R^2_{control}$，他们就能估计设计的功效。这一设计的效应值为：

$$f^2=\frac{R^2_{total}-R^2_{control}}{1-R^2_{total}} \qquad [10.7]$$

例如，假定研究者有 40 个观测值，全模型将解释因变量 30% 的变异（$R^2_{total}=0.3$），而包括 $c=4$ 个控制变量的模型仅能解释 20%的变异（$R^2_{control}=0.2$）。那么再增加 $k=2$ 个预测变量后的模型的功效是多少？F 检验在 $\alpha=0.05$ 水平上的临界值为 $F_{2,33}=3.285$。效应值为 $f^2=\frac{0.3-0.2}{1-0.3}=0.143$，从而非中心化参数为 $\lambda=0.143\times40=5.720$，因此相关的第二类错误为 0.478，功效为 0.522。

第 4 节 | 小结

本章为本书提供了一些背景脉络，告诉读者本书的目标以及为何聚焦于两组差异研究。本章还包含了有关其他类型分析的引用文献，这类分析往往需要更深入的功效分析。最后，本书的结尾部分讨论了观测数据的回归。



附　录

表 A.1 BMI 数据

	TREAT	BMI	PRE_BMI	TREAT	BMI	PRE_BMI
1	0	37.66	38.08	1	28.62	28.84
2	0	32.61	31.43	1	35.91	36.10
3	0	23.34	25.38	1	27.25	27.81
4	0	43.75	43.27	1	30.06	31.98
5	0	37.25	35.76	1	36.26	36.34
6	0	32.36	32.53	1	28.01	27.15
7	0	36.35	36.81	1	28.69	29.45
8	0	25.21	26.17	1	27.34	29.90
9	0	30.21	29.24	1	26.87	29.70
10	0	32.82	34.03	1	26.81	26.61
11	0	39.40	39.98	1	39.79	39.59
12	0	38.75	38.26	1	37.87	37.95
13	0	37.43	38.65	1	41.20	40.59
14	0	35.56	34.73	1	26.22	26.09
15	0	29.40	31.17	1	27.71	27.01
16	0	28.62	27.55	1	27.05	26.52
17	0	40.75	41.43	1	35.20	35.08
18	0	38.58	38.95	1	27.64	26.56
19	0	43.37	42.34	1	34.29	34.07
20	0	33.87	33.84	1	27.53	28.58
21	0	37.42	37.53	1	29.15	28.58
22	0	28.08	28.39	1	26.27	25.93
23	0	42.41	42.16	1	43.97	45.54
24	0	31.33	31.29	1	38.65	37.38
25	0	46.98	45.03	1	28.12	27.30

表 A.2 数学成绩数据

	SCHOOL	TREAT	MATH	PRE_MATH	MEAN_PRE_MATH
1	113	1	7	1	2.25
2	113	1	6	1	2.25
3	113	1	6	3	2.25
4	113	1	7	4	2.25
5	116	1	11	4	4.00
6	116	1	10	4	4.00
7	116	1	8	3	4.00
8	116	1	11	5	4.00
9	117	0	3	4	4.75
10	117	0	9	5	4.75
11	117	0	6	5	4.75
12	117	0	8	5	4.75
13	123	1	9	5	4.50
14	123	1	11	4	4.50
15	123	1	8	4	4.50
16	123	1	10	5	4.50
17	210	1	9	3	3.50
18	210	1	6	3	3.50
19	210	1	7	5	3.50
20	210	1	8	3	3.50
21	226	0	9	5	3.50
22	226	0	2	1	3.50
23	226	0	10	4	3.50
24	226	0	6	4	3.50
25	303	0	6	3	3.00
26	303	0	5	3	3.00
27	303	0	7	4	3.00
28	303	0	3	2	3.00
29	304	1	7	3	3.25
30	304	1	8	1	3.25
31	304	1	10	4	3.25
32	304	1	10	5	3.25
33	319	0	3	1	3.00
34	319	0	10	4	3.00
35	319	0	7	4	3.00
36	319	0	8	3	3.00
37	321	0	4	4	3.25
38	321	0	3	4	3.25
39	321	0	5	2	3.25
40	321	0	8	3	3.25

表 A.3 模拟多点随机试验数据

	K	Y	TREAT	X	MEAN_X
1	1	49.85	0	44.45	45.87
2	1	68.68	1	40.25	45.87
3	1	93.58	1	71.46	45.87
4	1	94.87	1	55.80	45.87
5	1	53.35	0	34.70	45.87
6	1	76.06	1	56.06	45.87
7	1	38.84	0	34.58	45.87
8	1	52.51	0	29.65	45.87
9	2	56.37	0	50.83	45.59
10	2	41.12	1	44.59	45.59
11	2	43.87	0	41.55	45.59
12	2	61.60	0	71.08	45.59
13	2	64.61	1	34.08	45.59
14	2	41.16	1	44.80	45.59
15	2	49.19	0	38.13	45.59
16	2	53.13	1	39.69	45.59
17	3	51.04	0	40.64	50.65
18	3	70.55	0	51.20	50.65
19	3	66.62	1	40.00	50.65
20	3	63.24	1	64.75	50.65
21	3	52.30	0	48.79	50.65
22	3	57.85	1	58.03	50.65
23	3	98.13	1	68.91	50.65
24	3	63.27	0	32.84	50.65
25	4	68.89	1	63.37	58.01
26	4	74.77	0	77.26	58.01
27	4	61.42	1	32.17	58.01
28	4	84.67	1	45.24	58.01
29	4	87.34	0	70.82	58.01
30	4	59.89	1	52.67	58.01
31	4	70.72	0	64.13	58.01
32	4	71.49	0	58.42	58.01
33	5	41.27	0	47.87	49.88
34	5	68.86	1	63.58	49.88
35	5	49.65	1	23.55	49.88
36	5	23.52	0	54.59	49.88
37	5	50.51	0	65.20	49.88
38	5	50.07	0	43.27	49.88
39	5	46.49	1	37.73	49.88
40	5	80.35	1	63.26	49.88

表 A.4 不同自由度的 t 分布分位数和标准正态分布分位数

df	分位数						
	$t_{(df)0.1}$	$t_{(df)0.2}$	$t_{(df)0.3}$	$t_{(df)0.95}$	$t_{(df)0.975}$	$t_{(df)0.99}$	$t_{(df)0.995}$
2	−1.886	−1.061	−0.617	2.92	4.303	6.965	9.925
3	−1.638	−0.978	−0.584	2.353	3.182	4.541	5.841
4	−1.533	−0.941	−0.569	2.132	2.776	3.747	4.604
5	−1.476	−0.92	−0.559	2.015	2.571	3.365	4.032
6	−1.44	−0.906	−0.553	1.943	2.447	3.143	3.707
7	−1.415	−0.896	−0.549	1.895	2.365	2.998	3.499
8	−1.397	−0.889	−0.546	1.86	2.306	2.896	3.355
9	−1.383	−0.883	−0.543	1.833	2.262	2.821	3.25
10	−1.372	−0.879	−0.542	1.812	2.228	2.764	3.169
11	−1.363	−0.876	−0.54	1.796	2.201	2.718	3.106
12	−1.356	−0.873	−0.539	1.782	2.179	2.681	3.055
13	−1.35	−0.87	−0.538	1.771	2.16	2.65	3.012
14	−1.345	−0.868	−0.537	1.761	2.145	2.624	2.977
15	−1.341	−0.866	−0.536	1.753	2.131	2.602	2.947
16	−1.337	−0.865	−0.535	1.746	2.12	2.583	2.921
17	−1.333	−0.863	−0.534	1.74	2.11	2.567	2.898
18	−1.33	−0.862	−0.534	1.734	2.101	2.552	2.878
19	−1.328	−0.861	−0.533	1.729	2.093	2.539	2.861
20	−1.325	−0.86	−0.533	1.725	2.086	2.528	2.845
21	−1.323	−0.859	−0.532	1.721	2.08	2.518	2.831
22	−1.321	−0.858	−0.532	1.717	2.074	2.508	2.819
23	−1.319	−0.858	−0.532	1.714	2.069	2.5	2.807
24	−1.318	−0.857	−0.531	1.711	2.064	2.492	2.797
25	−1.316	−0.856	−0.531	1.708	2.06	2.485	2.787
30	−1.31	−0.854	−0.53	1.697	2.042	2.457	2.75
35	−1.306	−0.852	−0.529	1.69	2.03	2.438	2.724
40	−1.303	−0.851	−0.529	1.684	2.021	2.423	2.704
45	−1.301	−0.85	−0.528	1.679	2.014	2.412	2.69
50	−1.299	−0.849	−0.528	1.676	2.009	2.403	2.678
55	−1.297	−0.848	−0.527	1.673	2.004	2.396	2.668
60	−1.296	−0.848	−0.527	1.671	2	2.39	2.66
65	−1.295	−0.847	−0.527	1.669	1.997	2.385	2.654
70	−1.294	−0.847	−0.527	1.667	1.994	2.381	2.648
75	−1.293	−0.846	−0.527	1.665	1.992	2.377	2.643
80	−1.292	−0.846	−0.526	1.664	1.99	2.374	2.639
85	−1.292	−0.846	−0.526	1.663	1.988	2.371	2.635
90	−1.291	−0.846	−0.526	1.662	1.987	2.368	2.632
95	−1.291	−0.845	−0.526	1.661	1.985	2.366	2.629
100	−1.29	−0.845	−0.526	1.66	1.984	2.364	2.626
	$z_{0.1}$	$z_{0.2}$	$z_{0.3}$	$z_{0.95}$	$z_{0.975}$	$z_{0.99}$	$z_{0.995}$
∞	−1.282	−0.842	−0.524	1.645	1.96	2.326	2.576

注释

[1] 譬如,两个类别变量的独立性检验就是 χ^2 检验。在大多数的导论性统计课上都会论及这一检验, $\sum_i \frac{(Q_i - E_i)^2}{E_i}$,其中 O_i 是观测频数,E_i 是期望频数。显然,我们是在计算与期望值距离的平方和。

[2] 自由度有时是个较难理解的概念。本质上,自由度是当我们已知某些取值的相关统计量后,这些取值能变动的个数。譬如,给定数值 5 和 7,如果我们已知均值是 6,那么在数据集{5, 7}中,我们其实只拥有一个信息。这是因为,如果我告诉你有两个数字,7 和一个未知数 a,但这两个数的均值是 6,那么我们就能通过式子计算出,$6=(7+a)/2$, $a=6\times2-7=5$。这就是为什么在处理列联表时,独立性检验的自由度等于行数目减去 1,然后乘以列数目减去 1。

[3] 如本书最后一章所示,F 分布也可用于相关和多元相关,因为相关系数是标准化的协方差。

[4] 如表 4.1 所示,样本的 BMI 均值为 33。BMI 取值 25 到 29.999 为超重,超过 30 为肥胖。

[5] 全称为政治和社会研究校际联盟(Inter-university Consortium for Political and Social Research),网址为 www.icpsr.umich.edu。

[6] 柯克(Kirk)与其他很多学者用 α_j 来表示干预效应,但在这本有关功效的概论中,第一类错误 α 扮演了极其重要的角色,我认为这种标记法有点混乱,因此我使用 τ_j。

[7] 为了避免与第二类错误参数 β 混淆,我用 γ 来标记回归斜率。

[8] 我用 T 而非 x 作为预测变量,是为了避免与后面将会论及的协变量相混淆。

[9] 后续统计检验的一个主要假定是控制组和干预组的方差相同。显然,在实际的数据中,这一假定不可能完全为真,但有许多检验可用于确定这一假定是否可能。我们对所使用的实例数据也进行了类似检验,结果表明数据中两个组的方差在统计上无差异。

[10] 需要注意的是,这个值略小于总标准差,即$\tilde{\sigma}^2 \neq \hat{\sigma}^2$。总标准差为 5.997,而合并标准差为 5.726。这是因为总标准差基于总均值计算而来,而合并标准差则基于各组均值计算而来(参见方程 4.4 和方程 4.5)。

[11] 由于四舍五入,这比表 5.1 中报告的均方误差值 5.727 略小。

[12] 由于四舍五入,这和计算机输出的结果 2.396 并不完全相等。

[13] 譬如,在 R 软件中这个函数是 pt,非中心化参数可以作为其中一个参

数输入。这一函数在 Stata 中被称为 nt(另有一个仅针对中心化分布的命令 t),在 SPSS 中则叫作 ncdf.t。每个程序都提供详尽的安装和在线帮助说明,用以解释输入参数的顺序。

[14] 即未引入干预的情况下,y 对 x 进行回归的总体斜率。

[15] 需要注意的是,就像合并方差那样,结果变量和协变量之间的总体相关估计值是基于对组均值的偏离得到的。换言之,这是未引入干预时 y 和 x 之间的相关。

[16] 其平方根为 -0.322,非常接近我们所观测的值,$\rho_{Tx}=-0.325$。

[17] 这并非进行功效分析的唯一方法。利普西(Lipsay, 1990)建议利用与干预指示变量不相关的协变量,使用$\dfrac{\bar{y}_1-\bar{y}_0}{\sqrt{\sigma^2(1-\rho^2_{yx,w})}}$来计算更大的效应值。在协变量不相关的情况下,该方法使研究者可以利用标准功效表。

[18] 由于分位数表中没有这一自由度,最好应使用计算机来求自由度 $2\times20-3=37$ 时的分位数,但本例中我们可以用一个近似值。

[19] 这意味着要用到 F 检验,所以读者需要回顾一下第 2 章的内容。

[20] 由于这是公众开放数据,所以原始数据已被处理成了等距间隔的定序类别。譬如,1 表示 11—15 分,2 表示 16—20 分,以此类推,最大的 11 表示 61 分及以上。所使用的协变量为一年级数学能力,编码为 1—5,1 表示处于最低的 10%,而 5 则处于最高的 90%—100%。

[21] 虽然不明显,但方程 6.6 可通过下列推导得到。若以$\bar{y}_1-\bar{y}$来估计 τ_j,则在两组的情况下,$n\sum_j\tau_j^2$ 的估计值为$n[(\bar{y}_1-\bar{y})^2+(\bar{y}_0-\bar{y})^2]$。而由于总均值的估计值为 $\bar{y}=(\bar{y}_1+\bar{y}_0)/2$,然后代入 $\hat{\tau}_1^2=(\bar{y}_1-\bar{y})^2$ 和 $\hat{\tau}_0^2=(\bar{y}_0-\bar{y})^2$ 进行运算,两者都等于$\dfrac{1}{4}(\bar{y}_1^2+\bar{y}_0^2-2\,\bar{y}_1\,\bar{y}_0)$。由于有两项 τ,则 $MSB=n\left[\dfrac{1}{2}(\bar{y}_1^2+\bar{y}_0^2-2\,\bar{y}_1\,\bar{y}_0)\right]$,而$\bar{y}_1^2+\bar{y}_0^2-2\,\bar{y}_1\,\bar{y}_0$ 即 $(\bar{y}_1-\bar{y}_0)^2$,所以得到 $MSB=\dfrac{n}{2}(\bar{y}_1-\bar{y}_0)^2$。

[22] 这里我还是不使用字母 β,以避免本书中的符号混乱。

[23] 然而,读者将会发现不同的单位和聚类数量组合会获得同样的功效,因此最主要的考量是成本(Raudenbush,1997)。

[24] 需要注意的是赫奇斯和罗兹(Rhoads)用了一个类似的比值 ω,其值为 υ 的一半,即 $\omega=\upsilon/2$。

[25] 这是因为很难拿到公开的多点试验数据。

[26] 本书未涉及抽样设计的其他要素,如抽样权重和分层抽样。建议读者

可在洛尔(Lohr，2009)的研究中寻找相关资源，关于抽样相关应用的更为技术性的讨论可参见沃尔特(Wolter，2007)。

[27] 这些“小”“中”和“大”效应值都是基于社会数据，甚或与特定干预的效应相关。

[28] 当然，这里的前提假定是某个干预的效应不可能等于或超过一年教育年限的效应。

[29] 若某项研究为复制性研究，这里提供的策略依然有效：如果研究未提供效应值，那就自己计算效应值。

[30] 据我所知，这方面的研究中还没有做过随机试验。

[31] 很明显，这不是一个正态分布变量，且分析应当是一个一般化(回归)模型。然而，我们通常很难找到一项完美的研究。文件检索中经常出现这类问题，所以这是非常现实的练习。

[32] 这里的分析单位是被访者，嵌套于警官。因此更为合适的分析应该考虑这一点。然而，即使研究者未考虑这一问题，未使用多层模型对效应值的影响也非常小。

[33] 如果两组标准差不同，我们就需要用各组样本量来计算合并标准差。

[34] 这无疑是一个随意的选择，但也是为了取整。

[35] 也存在其他的转化方式，譬如我们可以把发生比率(odds ratio)转化成相关系数。

[36] 无疑也存在其他类型的复杂样本，其中包括分层样本和分层聚类样本。很多复杂样本还使用概率权重来处理不均衡概率选择的问题。每个要素都会产生影响分析的一个设计效应(Lohr，2009)。分层会降低方差，而权重和聚类会增加方差。

[37] 这无疑是一个大城市。

[38] 实际上，用 F 分布可以进行很多检验(Murphey，Myors，& Wolach，2009)。

参考文献

Blalock, H. M. (1972). *Social statistics* (2nd ed.). New York, NY: McGraw-Hill.

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*, 547–556.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*. New York, NY: Russell Sage Foundation.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*, 43–82.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Brown, S. R., & Melamed, L. E. (1990). *Experimental design and analysis*. Thousand Oaks, CA: Sage.

Casella, G. (2008). *Statistical design*. New York, NY: Springer Science+Business Media.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. London, England: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Dahl, D. B. (2009). *xtable: Export tables to LaTeX or HTML (R package version)*.

Dhurandhar, E. J., Dawson, J., Alcorn, A., Larsen, L. H., Thomas, E. A., Cardel, M., ... Allison, D. B. (2014). The effectiveness of breakfast recommendations on weight loss: A randomized controlled trial. *American Journal of Clinical Nutrition*, *100*, 507–513.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, England: Arnold.

Easterbrook, P. J., Gopalan, R., Berlin, J., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet*, *337*, 867–872.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.

Gamoran, A., Turley, R. N. L., Turner, A., & Fish, R. (2012). Differences between Hispanic and non-Hispanic families in social capital and child development: First-year findings from an experimental study. *Research in Social Stratification and Mobility*, *30*, 97–112.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883.

Hedberg, E. C. (2012). *RDPOWER: Stata module to perform power calculations for random designs (Statistical Software Components)*. Chestnut Hill, MA: Boston College Department of Economics.

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*, 546–582.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research*. Washington, DC: National Center for Special Education Research.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, *17*, 1623–1634.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Jennings, W. G., Lynch, M. D., & Fridell, L. A. (2015). Evaluating the impact of police officer body-worn cameras (BWCs) on response-to-resistance and serious external complaints: Evidence from the Orlando Police Department (OPD) experience utilizing a randomized controlled experiment. *Journal of Criminal Justice*, *43*, 480–486.

Kirk, R. E. (1995). *Experimental design* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Lachin, J. M., & Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, *42*, 507–519.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, *55*(8), 1–24.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Liu, X. S. (2013). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. London, England: Routledge.

Lohr, S. L. (2009). *Sampling: Design and analysis*. Boston, MA: Cengage Learning.

Lohr, S. L. (2014). Design effects for a regression slope in a cluster sample. *Journal of Survey Statistics and Methodology*, *2*, 97–125.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

Mazerolle, L., Antrobus, E., Bennett, S., & Tyler, T. R. (2013). Shaping citizen perceptions of police legitimacy: A randomized field trial of procedural justice. *Criminology*, *51*, 33–63.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, England: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. Hoboken, NJ: Wiley.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*, 17–30.

Murphy, K., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. London, England: Routledge.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Murray, D. M., Rooney, B. L., Hannan, P. J., Peterson, A. V., Ary, D. V., Biglan, A., ... Schinke, S. P. (1994). Intraclass correlation among common measures of adolescent smoking: Estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology*, *140*, 1038–1050.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*, 492–510.

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, *34*, 383–392.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X.-F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199–213.

Rhoads, C. (2017). Coherent power analysis in multilevel studies using parameters from surveys. *Journal of Educational and Behavioral Statistics*, *42*, 166–194.

Rice, J. (2006). *Mathematical statistics and data analysis*. Boston, MA: Cengage Learning.

Ryan, T. P. (2013). *Sample size determination and power*. Hoboken, NJ: Wiley.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*, 238–266.

Schultz, E. (1955). Rules of thumb for determining expectations of mean squares in analysis of variance. *Biometrics*, *11*, 123–135.

Sharpsteen, C., & Bracken, C. (2013). *tikzdevice: R graphics output in LaTeX format. R package version 0.7.0.* Retrieved from http://CRAN.R-project.org/package=tikzDevice.

Spybrook, J. (2007). *Examining the experimental designs and statistical power of group randomized trials funded by the Institute of Education Sciences* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Spybrook, J., Raudenbush, S. W., Liu, X.-F., Congdon, R., & Martínez, A. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the "optimal design" software*. Ann Arbor: University of Michigan School of Education, Hierarchical Models Project.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*, 490–519.

White, M. D. (2014). *Police officer body-worn cameras: Assessing the evidence*. Washington, DC: Office of Community-Oriented Policing Services.

Wildt, A. R., & Ahtola, O. (1978). *Analysis of covariance*. Newbury Park, CA: Sage.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer Science+Business Media.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

译名对照表

alternative hypothesis	备择假设
ANCOVA(analysis of covariance)	协变量分析
ANOVA(analysis of variance)	方差分析
balanced design	均衡设计
cluster randomized trial	聚类随机试验
critical value	临界值
cumulative density function	累积密度函数
degree of freedom	自由度
density function	密度函数
design effect	设计效应
effect size	效应值
generalized linear model	一般线性模型
histogram	直方图
institutional review boards	伦理委员会/机构审查委员会
intraclass correlation	组内相关系数
margin of error	误差幅度
minimum detectable effect size	最低可检测效应值
MSE(mean squared error)	均方误差
multicolinearity	多重共线性
multisite randomized trials	多点随机试验
multiway ANOVA	多因素方差分析
noncentrality parameter	非中心化参数
null hypothesis	零假设/虚无假设
odds ratio	发生比率
one-tailedtest	单尾检验
ordinary least squares(OLS)	一般最小二乘法
precision analysis	精度分析
pooled variance/standard deviation	合并方差/标准差
power table	功效表
reliability	信度
reproducibility	可复制性

restricted maximum likelihood	限制性最大似然法
robustness	稳健性
sampling distribution	抽样分布
scale-free parameter	无标尺参数
shirt-sizes	“衬衣尺码”法
standard deviation	标准差
standard error	标准误
statistical power	统计功效
two-level cluster randomized trial	二层聚类随机试验
two-level multisite randomized trial	二层多点随机试验
two-tailed test	双尾检验
type I error	第一类错误
type II error	第二类错误
validity	效度
variance inflation factor(VIF)	方差膨胀因子



统计功效是指假定零假设为假、备择假设为真的情况下，统计检验显著的概率，是研究设计的关键组成部分。本书为读者提供了阅读复杂的功效分析技术论文和材料所需的背景、示例和解释，是一本清晰易懂的指南，阐释了测试统计数据的组成部分及其抽样分布。作者通过干预组和控制组两组间的比较，概述了功效分析的核心要素和机制。

主要特点

- 运用统计表达式的示例帮助读者理解复杂公式
- 通过演示设立好的研究假设，探讨功效分析文献的使用方法
- 介绍功效分析中必需的关键点，提供功效分析报告写作实例

上架建议：社会研究方法

ISBN 978-7-5432-3278-5



您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn



微信



天猫

定价：48.00元
格致网：www.hibooks.cn

[General Information]
书名=功效分析概论：两组差异研究
页数=204
SS号=15034139